Universidad Autónoma de Baja California

Instituto de Ingeniería

Maestría y Doctorado en Ciencias e Ingeniería

"*Evaluación de algoritmos para la inferencia de haplotipos, partiendo de genotipos tipo SNP"*

Tesis para obtener el grado de:

Maestro en Ciencias

Presenta:

Saúl Alejandro Roa Ledesma

Director:

Dr. Rafael Villa Angulo

Codirector:

Dr. Martín Luis Arango Pérez

Mexicali, B.C. Noviembre de 2022.

AGRADECIMIENTOS

A mi familia, hermanos, profesores y amistades que me apoyaron e inspiraron en cada momento de esta etapa tan anhelada en mi vida. Sus palabras de aliento, su comprensión, tiempo y retroalimentación me llenaron de confianza cuando las cosas se ponían muy difíciles.

Al Instituto de Ingeniería Mexicali de la Universidad Autónoma de Baja California y al Programa de Maestría y Doctorado en Ciencias e Ingeniería (MyDCI), por el apoyo, procuración y compromiso con el alumnado.

Agradecimientos especiales al Dr. Rafael Villa, siendo un ejemplo profesional y humano; le agradezco la oportunidad aceptándome en este proyecto. Su guía, sus comentarios y su paciencia hicieron este estudio posible.

Al Consejo Nacional de Ciencia y Tecnología (CONACYT) por otorgarme la beca para el financiamiento de mis estudios de maestría.

DEDICATORIA

A mis padres por siempre estar ahí, mi madre Norberta Ledesma, mi ejemplo de lucha y entrega, la persona más fuerte que conozco. Mi padre Rafael Roa†, su aliento a lo largo de mi vida académica perdura más allá, donde quiera que esté: “El estudio es todo, ¡órale!”.

Los tuve en mi mente en cada hora de trabajo.

“La ciencia es el trabajo de la mente humana, que está destinado más a estudiar qué a conocer, a buscar la verdad más que a encontrarla.”

Évariste Galois.

TABLA DE CONTENIDO

Página

ÍNDICE DE TABLAS

Página

ÍNDICE DE FIGURAS

Página

LISTA DE ABREVIATURAS

ADN	Ácido Desoxirribonucleico.
BAM	Archivo binario que se utiliza para representar secuencias genómicas alineadas.
BE	Extensión de bloque.
BIM	Archivo con información de las distancias genómicas de marcadores genéticos.
bp	Base par.
D'	Nivel de heredabilidad conjunta de dos SNPs
EM	Expectación Maximización.
FAM	Archivo con el formato de pedigrí.
G	Genotipo.
GWAS	Estudios de Asociación de todo el Genoma.
H	Haplotipo.
HI	Inferencia de Haplotipos.
HIPP	Inferencia de Haplotipos por Parsimonia Pura.
I	Individuo.
IBD	Idéntico por Descendencia.
ID	Número de Identificación.
ILP	Programación Lineal Entera
k-RHC	Configuración de Haplotipos con *k* Recombinaciones.
LD	Desequilibrio de ligamiento.
LNP	Ligamiento No Paramétrico.

MAP Archivo de mapa genético.

Mb Mega base par.

MCMC Método de Monte Carlo basado en Cadenas de Markov.

MLHC Configuración de Haplotipos de Máxima Probabilidad o Verosimilitud.

MRHC Configuración de Haplotipos con Mínimas Recombinaciones.

MRHCE Configuración de Haplotipos con Recombinaciones Mínimas con Errores Acotados.

PCR Reacción en Cadena de la Polimerasa.

PED Archivo de Pedigrí.

R Lenguaje de programación de código abierto, para análisis estadístico.

r^2 Correlación de Pearson

rs# Número de identificación, con respecto a las bases de datos de SNPs.

SAT Satisfactibilidad Booleana.

SNP Polimorfismo de Nucleótido Único o Simple.

UMD 3.1 Ensamblaje versión 3.1 del ganado Holstein.

VCF Variant Call Format, formato de almacenamiento de variaciones de secuencias.

WIDDE Web-Interfaced Next Generation Database dedicated to Genetic Diversity Exploration

wMEC Corrección Mínima de Errores Ponderada.

ZRHC Configuración de Haplotipos con Cero Recombinaciones.

Capítulo I

1.1 Introducción

A mediados de la década de los 70’s, el bioquímico británico Frederick Sanger desarrolló técnicas para la secuenciación del ADN, lo cual dio lugar al desarrollo de métodos automatizados de secuenciación y, por consiguiente, la comunidad científica pudo finalmente fijar el camino hacia uno de los proyectos más ambiciosos y profundos de la historia del ser humano: El Proyecto del Genoma Humano. Dicho proyecto consiste en un programa de investigación colaborativo a nivel internacional para el ensamblaje y entendimiento completo de todos los genes del ser humano, iniciando en el año 1990, y para febrero del año 2001 ya se contaba con una versión preliminar del 90% de nuestro genoma [1].

Este proyecto promovió el desarrollo de nuevas y mejores tecnologías de secuenciación, creando así enormes cantidades de datos de genotipos con un coste económico cada vez menor. Así pues, los estudios de asociación de todo el genoma (GWAS) presentan muchos nuevos retos computacionales y estadísticos, debido a la gran cantidad de datos que se obtienen actualmente.

Un proyecto al respecto surge en el 2013 mediante la SAGARPA (Secretaría de Agricultura, Ganadería, Desarrollo Rural, Pesca y Alimentación), en conjunto con la CONARGEN (Consejo Nacional de los Recursos Genéticos Pecuarios), para el mejoramiento de diversas razas comerciales de ganado en nuestro país, con el objetivo de posicionar la genética mexicana, en América Latina, abrir nuevos nichos de mercado y consolidar los existentes.

Aunque algunos estudios de asociación de enfermedades pueden realizarse utilizando únicamente alelos de un solo locus o frecuencias de genotipos, la información sobre los haplotipos es esencial para el análisis detallado de los mecanismos de una enfermedad. La identificación de los haplotipos permite realizar pruebas de asociación con enfermedades basadas en los haplotipos. Esto es especialmente importante en los estudios de asociación de todo el genoma. De hecho, los estudios de asociación haplotípica han encontrado posiciones de ADN asociadas a enfermedades que no son significativas en todo el genoma utilizando pruebas de un solo marcador. Además, la mayoría de los métodos de imputación requieren datos haplotípicos en lugar de genotipos [2].

1.1.1 Objetivo general

Analizar el funcionamiento de distintos algoritmos presentes en el estado del arte para la inferencia de haplotipos a partir de genotipos tipo SNPs en un pedigrí promedio de ganado bovino.

Recomendar, basado en los resultados obtenidos, el mejor algoritmo para inferencia de haplotipos en el ganado Mexicano analizado por la Comisión Nacional de Reserva Genética, en un proyecto de genotipificación de 10,000 cabezas de ganado bovino Mexicano.

1.1.2 Objetivos específicos

Los objetivos específicos son los siguientes:

- Realizar un estudio del estado del arte en algoritmos para la inferencia de haplotipos en pedigrís, partiendo de genotipos SNP.
- Generar un concentrado de genotipos SNP para el desarrollo del proyecto.

- Evaluar los algoritmos en cuanto a precisión, tiempo de ejecución y número de recombinaciones de los algoritmos existentes para la inferencia de haplotipos, en un conjunto de individuos de ganado bovino relacionado genéticamente en un árbol genealógico (pedigrí) promedio con una estructura compleja, a lo largo de un número de marcadores de 5, 10, 20 y 50 SNPs.

1.1.3 Polimorfismos de Nucleótido Único (SNPs)

Uno de los primeros pasos a seguir para el estudio y comprensión de enfermedades complejas, y la aparición de rasgos fenotípicos de interés en mamíferos mayores, es la genotipificación de la especie en cuestión, aunado al análisis de las diferencias estructurales y puntuales que existen entre el genotipo de las poblaciones estudiadas.

Una de las herramientas utilizadas en los estudios de asociación es la obtención de una clase de marcador genético ampliamente utilizado por su abundancia, y porque éstos representan la clase más común de polimorfismos, estos son los llamados Polimorfismos de Nucleótido Único (SNP), ilustrados en la figura 1.

Los SNPs son variaciones de secuencia de ADN que se producen cuando se altera un solo nucleótido: adenina (A), timina (T), citosina (C) o guanina (G) en la secuencia del genoma y se distinguen de las mutaciones mediante su frecuencia de aparición, es decir, se observan con mayor frecuencia en la población general que las mutaciones.

Representan a las variantes genéticas más comúnmente encontradas en el genoma. Debido a su amplia distribución, estos polimorfismos se localizan en cualquier parte de la estructura de los genes y el genoma [3].

Figura 1. "Polimorfismo de nucleótido simple (SNP)."

1.1.4 SNPchip

Son chips de ADN, de oligonucleótidos inmovilizados de secuencias conocidas, que difieren en sitios específicos de nucleótidos individuales (en el sitio del SNP). La técnica es adecuada para interrogar varios SNPs en paralelo de cada muestra de una manera multiplexada. Hace uso de la técnica de secuenciación mediante hibridación. Cuatro oligonucleótidos en una columna de una matriz difieren sólo en el sitio SNP y sólo uno sería totalmente homólogo, como se muestra en la figura 2. Posteriormente una matriz de este tipo se hibrida con el producto de la PCR [4].

Los microarreglos (laminillas) de SNP tienen tres componentes importantes:

1. Secuencias de ácido nucleicos objetivo, inmovilizadas en portaobjetos de vidrio (Affymetrix) o microperlas (Infinium);

2. Una o más sondas oligonucleótidas marcadas específicas para cada alelo; y

3. Un sistema de detección para registrar y "traducir" la señal de hibridación.

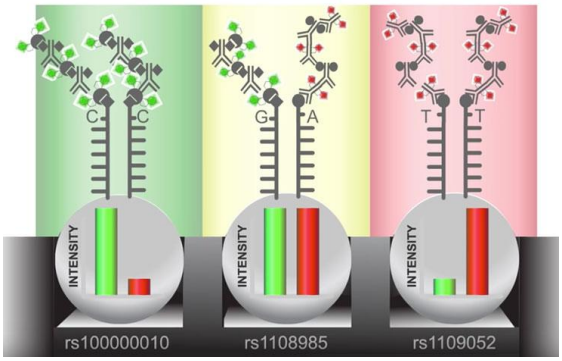


Figura 2. "Ilustración del funcionamiento de los SNPChips Illumina."

A medida que los fragmentos de ADN pasan por el chip, cada sonda se une a una secuencia complementaria en el ADN de la muestra, deteniéndose una base antes del locus de interés. La especificidad alélica se confiere mediante una extensión de una sola base que incorpora uno de los cuatro nucleótidos marcados. Al ser excitado por un láser, el nucleótido marcado emite una señal. La intensidad de esa señal transmite información sobre la proporción alélica en ese locus [5].

1.1.5 Aplicaciones de la genotipificación de SNPs

Las aplicaciones de esta tecnología se han expandido considerablemente en distintas direcciones, incluyendo nuevas aplicaciones clínicas, pero también representan nuevas herramientas en áreas tan diversas como la ecología, la genética de poblaciones, la evolución y el medio ambiente, u otras aplicaciones donde se compara la lista de transcriptomas de dos individuos de la misma especie.

La genotipificación de SNP de alta densidad o la secuenciación del genoma están disponibles como diagnóstico biomédico para predecir la predisposición individual a las enfermedades genéticas hereditarias, marcando el comienzo de la era de la "genómica personal". Aplicaciones similares de

estas tecnologías en especies ganaderas como el ganado vacuno tienen como objetivo la mejora de la productividad y la salud de los animales, así como la precisión de la selección dentro de los programas de mejora genética. En el ganado, dichos estudios se pueden utilizar para localizar las regiones genómicas que contribuyen a la variación genética natural en cualquier rasgo fenotípico de interés [6].

1.2 Inferencia de Haplotipos (HI)

Los humanos somos diploides, con dos cromosomas homólogos, uno de cada padre. Las tecnologías actuales de genotipificación producen genotipos en desfase, es decir, pares de alelos desordenados, con orígenes parentales no identificados. Un haplotipo se refiere a una combinación de alelos en un solo cromosoma, con la característica de que todos los alelos tienen el mismo origen parental.

El problema de la inferencia de haplotipos consiste en determinar dicho origen parental de los alelos y, en consecuencia, sus haplotipos a partir de genotipos desfasados utilizando métodos experimentales o computacionales.

Los haplotipos confieren información fundamental para comprender la genética de las enfermedades. Sin embargo, la información sobre los haplotipos no puede obtenerse directamente de las técnicas experimentales de genotipificación actuales [7].

La determinación de la fase de haplotipos es cada vez más importante a medida que entramos en la era de la secuenciación a gran escala porque muchas de sus aplicaciones, como imputar variantes de baja frecuencia y caracterizar la relación entre la variación genética y la susceptibilidad a la enfermedad, son particularmente relevantes [8].

1.3 Recombinación genética

En la profase de la meiosis, los cromosomas homólogos se emparejan y parecen mantenerse unidos principalmente en los quiasmas, los lugares donde sus cromátidas se enroscan entre sí. A veces, en un quiasma, las cromátidas se rompen físicamente y luego intercambian segmentos, en un proceso llamado entrecruzamiento. Si se produce un cruce entre dos loci, sus alelos serán reordenados, como se muestra en la figura 3.

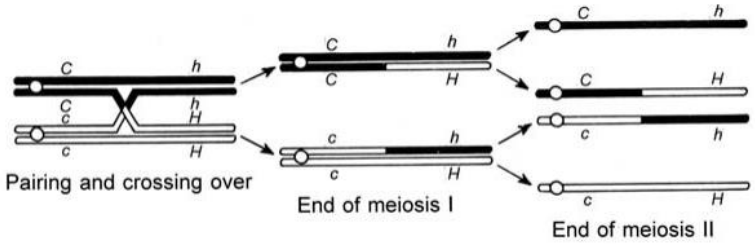


Figura 3. "Proceso de recombinación en la meiosis."

El cruce produce recombinación, y la frecuencia de recombinación, denotada por *R*, es el número de recombinantes dividido por el número total de descendientes. Dichos cruces se producen al azar y dependen de la distancia entre los genes, si la distancia entre dos puntos es muy corta, hay poca probabilidad de que se produzca un cruce allí, pero si la distancia es mayor, se producirán cruces en esos puntos con mayor frecuencia, produciendo más recombinantes [9].

Capítulo II

2.1 Planteamiento matemático del problema de la Inferencia de Haplotipos (HI)

Formalmente, un haplotipo puede describirse como una cadena binaria, donde el valor 0 representa el nucleótido de tipo silvestre (el encontrado de manera natural y por lo general en mayor frecuencia dentro de cierta población) y el valor 1 representa el nucleótido de tipo mutante.

Los genotipos pueden describirse como cadenas sobre el alfabeto $\{0, 1, 2\}$. Cada SNP (también llamado sitio o posición) en un genotipo g_i de tamaño m se representa por $g_{i,j}$, con $1 \leq j \leq m$. Un sitio $g_{i,j}$ es homocigótico si $g_{i,j} = 0$ o $g_{i,j} = 1$. En caso contrario, cuando $g_{i,j} = 2$, el sitio es heterocigoto.

Por ejemplo:

Considere que el genotipo 02212 tiene 5 sitios, de los cuales un sitio es homocigoto con valor 0, un sitio es homocigoto con valor 1 y los tres sitios restantes corresponden a sitios heterocigotos. Hay cuatro posibles explicaciones para este genotipo: (00010, 01111), (00110, 01011), (00111, 01010) y (00011, 01110) [10].

2.2 Aplicaciones de la HI

Las aplicaciones de la inferencia de haplotipos incluyen: la comprensión de la interacción de la variación genética y la enfermedad, la imputación de la variación genética, la determinación de genotipos en datos de microarreglos y de secuenciación, la detección de errores de

genotipificación, estudios de histocompatibilidad, la inferencia de la historia demográfica, la inferencia de puntos de recombinación, la detección de mutaciones recurrentes y firmas de selección, y el modelado de la regulación de la expresión genética [8].

En el ámbito de la investigación, los haplotipos se utilizan habitualmente para localizar un gen o locus causante de la enfermedad. En la actualidad, el uso de los estudios de asociación genética suscita un gran interés, ya que se considera que este diseño de estudio es más potente que los estudios de vinculación a la hora de localizar los loci de susceptibilidad para enfermedades comunes [11].

2.2.1 GWAS

Un estudio de asociación de todo el genoma (GWAS) es un enfoque a gran escala para asociar variantes genéticas con resultados observables (fenotipos) en una población. Los estudios de asociación proceden mediante la identificación de un número de individuos que presentan una enfermedad o rasgo, y compara a estos individuos con aquellos que no lo portan, o no se sabe que lo portan. Ambos conjuntos de individuos son entonces genotipificados para un gran número de variantes genéticas (SNP), que luego se prueban para la asociación a la enfermedad/rasgo. Estos estudios han sido capaces de identificar con éxito un gran número de polimorfismos asociados a la enfermedad. Los estudios con decenas de miles de individuos se están volviendo comunes y son cada vez más el estándar en la asociación de variantes genéticas a la enfermedad [12].

Los GWAS suelen implicar la genotipificación directa de varios cientos de miles de SNP en cientos o miles de muestras de ADN mediante la tecnología de microarreglos. Tras estrictos procedimientos de control de calidad, cada variante se analiza en relación con el rasgo de interés.

Normalmente, los investigadores colaboran y combinan los datos de los estudios con la misma enfermedad o rasgo disponible [13].

2.3 Estrategias computacionales para la Inferencia de Haplotipos (HI)

Un método de reconstrucción de haplotipo se basa en los dos componentes siguientes:

1. El modelo genético reagrupa suposiciones realistas sobre el patrón de haplotipos que se espera en una población.
2. El algoritmo computacional determina cuál, entre todas las reconstrucciones de haplotipo candidatas de la muestra, es la más consistente con el modelo genético.

Existen tres tipos de enfoques computacionales para la HI:

1. Algoritmos combinacionales: Los algoritmos combinatorios consideran a su vez cada posible reconstrucción de haplotipo de los genotipos y luego eligen el más realista según una función de puntuación.
2. Algoritmos estadísticos: Los algoritmos estadísticos consideran la reconstrucción del haplotipo como un conjunto de parámetros desconocidos cuyos valores deben estimarse dados los genotipos observados.
3. Los algoritmos bayesianos consideran la reconstrucción del haplotipo como un conjunto de variables aleatorias discretas y estiman su distribución conjunta dados los datos de genotipo observados y las suposiciones previas sobre la distribución de haplotipos [2].

Las tecnologías de secuenciación genética de siguiente generación han permitido el desarrollo y estudio de los genotipos y variaciones genéticas de individuos no relacionados, así como de familias, como se verá en los siguientes apartados.

En teoría debería haber cuatro tipos de alelos en un sitio del SNP, pero en realidad, la mayoría de los SNP son bialélicos donde sólo aparecen dos tipos de alelos, que pueden ser simplemente denotados por '0' y '1' [14].

Existen 5 tipos de algoritmos estadísticos y computacionales para la solución del problema de inferencia de haplotipos:

1. Parsimonia
2. Filogénicos
3. Máxima verosimilitud
4. Inferencia Bayesiana
5. Algoritmos Genéticos

2.3.1 Métodos basados en parsimonia

Es un enfoque combinatorio, que está indirectamente relacionado con el modelo coalescente [15]. La inferencia de haplotipos por parsimonia pura (HIPP) tiene como objetivo encontrar un conjunto de haplotipos *H* que pueda explicar un conjunto dado de genotipos *G*. La motivación para buscar una solución de inferencia de haplotipos con el menor número de haplotipos está motivada biológicamente por el hecho de que los individuos de una misma población tienen los mismos ancestros y las mutaciones no se producen con frecuencia. Además, es un hecho bien conocido que el número de haplotipos en una población es mucho menor que el número de genotipos. Además, los resultados experimentales apoyan que las iteraciones del método de Clark devuelven soluciones más precisas cuando el número de haplotipos es menor, método basado en parsimonia que se discutirá en la siguiente sección.

2.3.2 Algoritmo de Clark

Clark, fue el primero en proponer un algoritmo para resolver el problema de inferencia de haplotipos (ver figura 4). Este algoritmo ha sido ampliamente utilizado, y todavía se utiliza hoy en día.

De forma abstracta, la entrada consiste en *n* vectores, cada uno de ellos de longitud *m*, donde cada valor en el vector es un genotipo que contiene: 0, 1 o 2. Cada posición de un vector está asociada a un sitio de interés en el cromosoma. El estado de cualquier sitio en el cromosoma es 0 y 1. La posición asociada en el vector tiene un valor de 0 o 1 si el sitio del cromosoma tiene ese estado en ambas copias (es un sitio homocigoto), y tiene un valor de 2 si ambos estados están presentes (es un sitio heterocigoto). Una posición se considera "resuelta" si contiene 0 o 1, y "ambigua" si contiene un 2. Un vector sin posiciones ambiguas se denomina "resuelto", y en caso contrario se denomina par de haplotipos "ambiguo" [10].

Se inicializa un conjunto de haplotipos H, mediante la identificación de todos los genotipos resueltos en el conjunto de genotipos, G. basados en el conjunto inicial H, los genotipos residuales (ambiguos) en G son resueltos uno a uno, y se van añadiendo a H, de acuerdo con las soluciones de estos haplotipos. Este proceso continúa hasta que no se puede resolver ningún genotipo [14].

Figura 4. "Aplicación de la regla de inferencia de Clark. Se muestra el conjunto de genotipos inicial, el inicial resuelto y el conjunto inicial ambiguo."

Gusfield presentó un modelo completamente diferente llamado modelo coalescente. La historia evolutiva se considera un árbol, con cada hoja del árbol representando un haplotipo. Este árbol debe ser generado por los haplotipos en el conjunto de resultados H. El modelo coalescente limita aún más las soluciones a los haplotipos y facilita este problema. La inferencia de haplotipos basada en este modelo también se llamó inferencia de haplotipos basada en filogenia perfecta [16].

2.3.3 Métodos Filogénicos

Están basados en el modelo biológico de coalescencia. La observación clave es que, en ausencia de recombinación, cada secuencia tiene un único ancestro en la generación anterior, ejemplo ilustrado en la figura 5. Es decir, si seguimos hacia atrás en el tiempo la historia de un único haplotipo H de un determinado individuo I, cuando no hay recombinaciones, ese haplotipo H es una copia de uno de los haplotipos de uno de los padres del individuo I. No importa que tenga dos

padres o que cada uno de ellos tenga dos haplotipos. La historia hacia atrás de un solo haplotipo en un solo individuo es un camino simple, si no hay recombinación.

Hay un elemento adicional del modelo básico de coalescencia: la suposición de infinitos sitios. Es decir, los m sitios de la secuencia (sitios SNP en nuestro caso) son tan escasos en relación con la tasa de mutación, que en el marco temporal de interés habrá ocurrido como máximo una mutación (cambio de estado) en cualquier sitio [16]. La figura 5 muestra un ejemplo de este modelo:

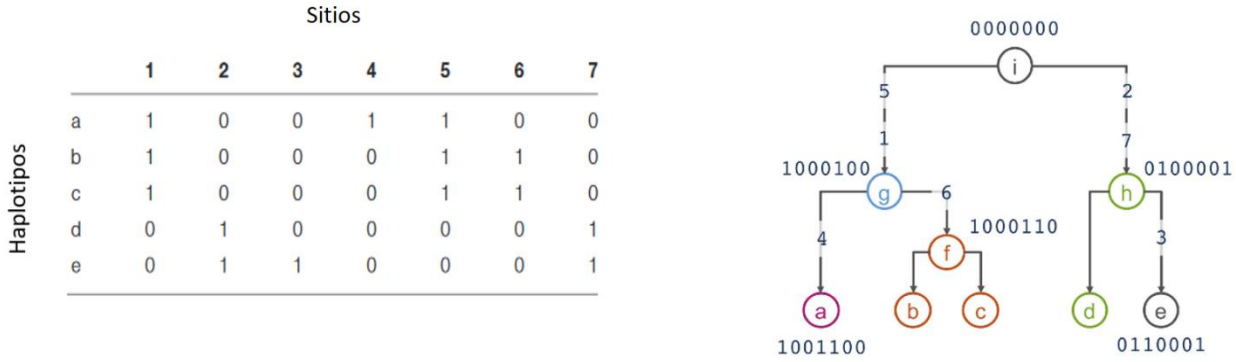


Figura 5. "Representación de Haplotipos en matriz binaria, y como filogenia perfecta".

En la figura 5, se asume que haplotipos similares tendrán efectos similares al ser heredados, por lo que es necesario modelar dichos efectos. Esto puede hacerse como sigue:

$$
\begin{aligned}
h_i &\sim N\left(0, \sigma^2_{h_m}\right) \\
h_{g'}|h_i &\sim N\left(\rho h_i, \sigma^2_{h_c}\right) \\
h_g|h_{g'} &\sim N(\rho h_{g'}, \sigma^2_{h_c}) \\
h_a|h_g &\sim N\left(\rho h_g, \sigma^2_{h_c}\right) \\
h_f, h_b, h_c, |h_g &\sim N\left(\rho h_g, \sigma^2_{h_c}\right) \\
h_{h'}|h_i &\sim N\left(\rho h_i, \sigma^2_{h_c}\right) \\
h_h, h_d|h_{h'} &\sim N\left(\rho h_{h'}, \sigma^2_{h_c}\right) \\
h_e|h_h &\sim N\left(\rho h_h, \sigma^2_{h_c}\right)
\end{aligned}
$$

Donde h_i, h_g, . . . , h_e, indican el efecto de los haplotipos *i*, *g*, . . . , *e*; y h_*‘ indica el efecto de haplotipos generados cuando suceden mutaciones. Por ejemplo: *g’* es el haplotipo generado entre los haplotipos *i* y *g*, debido a mutaciones entre *i* y g.

Dados los efectos anteriores, para construir el modelo es necesario considerar la densidad conjunta de todos los efectos de los haplotipos; la cual es expresada como sigue:

$$\boldsymbol{h}|\rho,\sigma^2_{h_c} \sim \mathcal{N}(\mathbf{0}, \mathbf{V}_h\left(\rho,\sigma^2_{h_c}\right))$$

lo cual produce la siguiente función de masa de probabilidad:

$$p(\boldsymbol{h}|\rho,\sigma^2_{h_c}) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \sigma^{-n}_{h_c}(1-\rho^2)^{1/2} \exp\left(-\frac{1}{2\sigma^2_{h_c}}\boldsymbol{h}^T\mathbf{H}\,(\rho)^{-1}\boldsymbol{h}\right).$$

Donde $\boldsymbol{V_h}$ es la covarianza de los efectos de los haplotipos, y $\boldsymbol{H}$ es una matriz de coeficiente de covarianzas.

2.3.4 Métodos basados en Máxima Verosimilitud

El problema de configuración de haplotipos de máxima verosimilitud (MLHC) tiene como objetivo encontrar la solución de haplotipos para todos los miembros del pedigrí que maximice la probabilidad de observar los genotipos dados. Existen varios métodos que resuelven exactamente el planteamiento del MLHC, entre ellos *Genehunter* [17], *Allegro* [18] y *Merlín* [19], y hay otros que realizan algoritmos aproximados, por ejemplo, *SimWalk2* [20]. Para la resolución de haplotipos, este método es implementado como sigue:

El propósito es estimar la frecuencia λ de los haplotipos en la población, basado en su verosimilitud, *L*, dado el conjunto de genotipos *D*. *L* es definido como la probabilidad de los genotipos contenidos en *D*, como sigue:

$$L = P_r(D \mid \lambda) \approx \prod_{i=1}^{n'} P_{r\lambda}(g_i)^{f_i} = \prod_{i=1}^{n'} \Big(\sum_{\{\forall <h_k, h_l> \mid h_k \oplus h_l = g_i\}} P_{r\lambda}(h_k, h_l) \Big)^{f_i}$$

Donde:

L = verosimilitud de los datos D,

g_i = genotipo i,

f_i = frecuencia de g_i en D,

$p_{r\lambda}(g_i)$ = probabilidad de g_i ,

$p_{r\lambda}(h_k, h_i)$ = probabilidad conjunta de todos los pares de haplotipos que resuelven el genotipo g_i..

Una alternativa para estimar las frecuencias λ de los haplotipos, dadas las frecuencias observadas de los genotipos, es aplicar el método de Expectación-Maximización (E-M). Este es un método iterativo para estimar conjuntos de frecuencias de haplotipos $p_1, p_2, p_3, \ldots, p_h$; partiendo de valores iniciales arbitrarios $p_1^{(0)}, p_2^{(0)}, p_3^{(0)}, \ldots, p_h^{(0)}$. Estos valores iniciales son utilizados para estimar frecuencias de pares de haplotipos $\tilde{p}(h_k h_l)$, y con ellas las frecuencias de los genotipos $p_j^{(g)}$ como si ellas fueran frecuencias desconocidas (***Paso de expectación***). Las frecuencias estimadas de pares de haplotipos son estandarizadas y usadas para estimar frecuencias de haplotipos $\hat{p}$ en la siguiente iteración (***Paso de Maximización***). La formulación matemática del procedimiento E-M es la siguiente:

- ✓ **Pasó de Expectación:**

$$\widetilde{P}\left(h_k h_l\right)^{(g)} = \begin{cases} p_k^{(g)2} & si\ k = l, \\ 2p_k^{(g)} p_l^{(g)} & si\ k \neq l. \end{cases}$$

- ✓ **Paso de Maximización:**

1. $$P_j^{(g)} = \sum_{i=1}^{c_j} P(genotipo\_i)^{(g)} = \sum_{i=1}^{c_j} \widetilde{P}(h_k h_l)^{(g)}$$
2. $$P(h_k h_l)^{(g)} = \frac{n_j}{n} \frac{\widetilde{P}(h_k h_l)^{(g)}}{P_j^{(g)}}$$
3. $$\hat{p}_t^{(g+1)} = \frac{1}{2} \sum_{j=1}^{m} \sum_{i=1}^{c_j} \delta_{it} P_j (h_k h_l)^{(g)}$$

Donde: δ_{it} es una variable que indica el número de veces que el haplotipo *t* está presente en el genotipo *i* (0, 1 o 2.)

2.3.5 Métodos Bayesianos.

Los algoritmos bayesianos tienen como objetivo estimar la distribución posterior de los parámetros a partir de los datos observados, asumiendo cierto conocimiento previo sobre la distribución de los parámetros. La distribución posterior puede estimarse utilizando, por ejemplo, técnicas de muestreo de Gibbs. El muestreo de Gibbs es un tipo de procedimiento de Monte Carlo con cadena de Markov. El objetivo es extraer muestras de una distribución posterior. El muestreo de Gibbs construye una cadena de Markov cuya distribución estacionaria es la verdadera distribución conjunta [10]. Una forma de implementar esta estrategia es la siguiente:

Las frecuencias de los haplotipos son asumidas como cantidades aleatorias no observadas y se evalúa su distribución condicional en función de los genotipos existentes, como sigue:

Para la distribución a priori:

$G = (G_1 , . . ., G_n)$; genotipos observados en los individuos de la población, en L locus.

$H = (H_1 , . . ., H_n)$; haplotipos actuales (no observados).

$r = (r_1 , . . ., r_n)$; vector de recombinaciones (desconocido), entre cada par de locus.

Hace uso del algoritmo de Montecarlo con Cadenas de Markov (MCMC) para estimar H partiendo de los valores observados G. Toma en cuenta el desequilibrio por ligamento y la taza de recombinación. Estima los haplotipos muestreando de $P_r = (H,\rho\backslash G)$.

2.3.6 Algoritmos Genéticos

Son algoritmos que codifican soluciones en vectores de 1's y 0's, llamados "cromosomas", que aplican operaciones inspiradas en la genética, tales como entrecruzamiento, mutación, e inversión. Dichos cromosomas se heredan de una población a otra mediante "selección natural". Dicha selección se lleva a cabo optimizando cierta función Aptitud, en relación a los parámetros de importancia en el problema a resolver. Después de determinadas épocas o iteraciones, se selecciona la solución o soluciones optimas que más se ajusten a la función Aptitud, siendo éstas las soluciones arrojadas por el programa [21]. El proceso básico de un algoritmo genético aplicado a resolver el problema de la inferencia de haplotipos se describe de la siguiente manera:

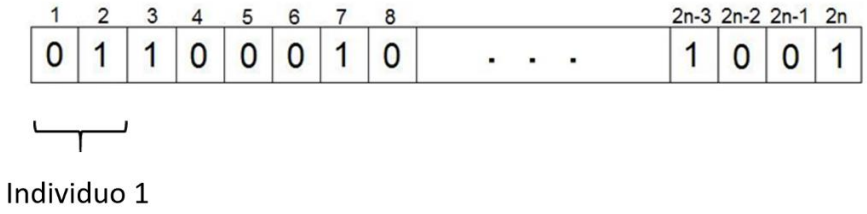
1. Generación de poblaciones aleatorias de n cromosomas (haplotipos candidatos válidos).
2. Evaluación de la Aptitud de cada cromosoma (número de recombinaciones).
3. Se seleccionan dos cromosomas como padres, de acuerdo a su Aptitud.

4. Mediante una probabilidad de cruzamiento P_c, se recombinan estos padres para formar nueva descendencia.
5. Con una mutación P_m, se inducen mutaciones en la descendencia.
6. Se añade la nueva descendencia a la nueva población.
7. Si no se satisface la condición, se vuelve al paso 2.
8. Si se cumple, arroja la solución y termina el proceso [22].

Los tres aspectos más importantes e influyentes para el éxito de una Algoritmo Genético son: 1) hacer una representación adecuada y precisa del espacio de búsqueda de soluciones; 2) la forma en que es representada una solución candidata, y; 3) la función aptitud que es evaluada para encontrar la solución óptima final. A continuación, se representa una forma válida de implementar estos tres aspectos para inferencia de haplotipos en pedigrís.

2.3.6.1 El espacio de búsqueda

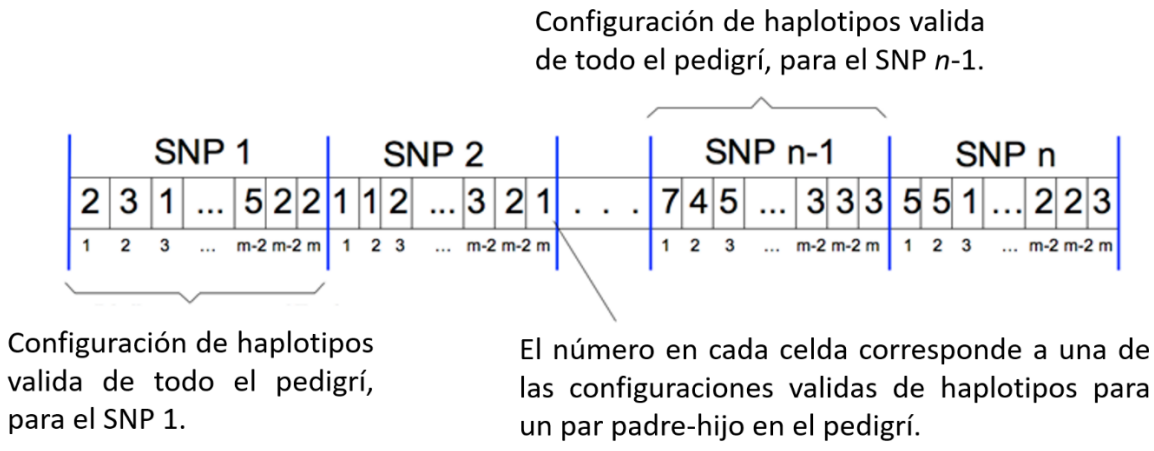
El espacio de búsqueda corresponde al número total de configuraciones validas de haplotipos para el pedigrí completo. Ahora, si representamos un SNP para todo el pedigrí como un vector donde cada dos celdas corresponden a los dos valores heredados por los padres a un individuo, y cada valor puede ser un 0 si el valor es heredado por el padre, o 1 si es heredado por la madre; entonces para cada SNP el especio de posibles combinaciones, que corresponde al espacio de búsqueda del SNP, estará dado por 2^{2n}.



Para obtener el tamaño del espacio total de todas las soluciones, incrementamos el espacio de búsqueda para el total de SNPs en la muestra. Suponiendo que se tienen *m* SNPS, entonces el espacio total de búsqueda será de: $m*2^{2n}$.

2.3.6.2 Representación

Para la inferencia de haplotipos en pedigrís, los datos iniciales corresponden a un conjunto de marcadores SNPs que fueron genotipificados para cada individuo en el pedigrí. Para construir una solución, es generado para cada SNP un vector que representa una solución válida de haplotipos para cada par padre-hijo en el pedigrí. Ahora, cada par padre-hijo tiene distintas configuraciones de haplotipos validas, las cuales son enumeradas. De esto que un SNP será un vector de números enteros representando en cada celda una configuración válida para cada par padre-hijo. Para generar la representación total de todo el pedigrí, con todos los SNPs; se genera un vector yuxtaponiendo los vectores de cada SNP. El vector resultante será la representación de una solución válida para todo el pedigrí con todos los SNPs. En este caso, cada SNP será tomado como un gen, y todo el conjunto de SNPs será un cromosoma completo. Un ejemplo se muestra a continuación:



2.3.6.3 Función aptitud

Para saber que tan bien un cromosoma (solución válida) resuelve el problema de inferencia de haplotipos en el pedigrí, es definida una función aptitud que asigna puntuación a cada solución, basada en el parámetro que se desee optimizar. Un parámetro comúnmente utilizado para evaluar es el número de recombinaciones, por lo que la función aptitud evalúa el número de recombinaciones que contienes cada solución valida que genera el algoritmo, como sigue:

$$Aptitud = 1 - \frac{Número\ calculado\ de\ recombinaciones}{Número\ maximo\ de\ recombinaciones}$$

Donde el máximo número de recombinaciones es igual a: (número de SNPs -1)*(2*número de individuos).

2.4 Inferencia de Haplotipos en pedigrís

Formalmente, un pedigrí es un grafo acíclico dirigido $G = (V, E)$, donde $V = M \cup F$ y M representa los nodos masculinos y F los femeninos. El grado interno de cada nodo es 0 o 2. Los nodos con grado 0 se llaman fundadores, mientras que los nodos con grado 2 se llaman no fundadores. En los nodos no fundadores, una arista entrante debe comenzar en un nodo masculino (llamado padre) y la otra arista entrante debe comenzar en un nodo femenino (llamado madre) y el nodo en sí se llama hijo.

Un subgrafo constituido por una madre, un padre y un hijo respectivo es un trío. Un pedigrí tiene un bucle de apareamiento si hay dos caminos diferentes desde un nodo x a un nodo y. Según

las leyes mendelianas de la herencia, cada sitio de un mismo haplotipo se hereda del mismo padre, suponiendo que no haya mutaciones dentro de un pedigrí. Por tanto, un cromosoma procede de la madre, mientras que el otro cromosoma procede del padre.

Descripción del problema: Dado un conjunto de n genotipos, G, cada uno de tamaño m, organizados en p pedigrís, el problema consiste en obtener el conjunto de haplotipos H, que expliquen el genotipo G y asocie un par de haplotipos $(H_i^a,\ H_i^b)$, donde $\left(H_i^a,\ H_i^b\right) \in H$ pertenecen a H, para cada genotipo $g_i \in G$, que satisfagan las leyes mendelianas de herencia y asumiendo que no hay mutaciones en los pedigrís [23]. En comparación con la inferencia del haplotipo en una población con individuos no relacionados, la estructura del pedigrí puede proporcionarnos mucha más información [14].

Existen un gran número de algoritmos para inferir haplotipos para individuos no relacionados, para tríos, y para dúos, sin embargo, en el caso de los algoritmos para la inferencia de haplotipos en pedigrís, este número decae, conforme a la complejidad del pedigrí a analizar aumenta, es decir, algoritmos para inferir haplotipos en pedigrís grandes, con recombinaciones, con lazos de endogamia, y con genotipos faltantes, son realmente escasos. La tabla 1 presenta los principales algoritmos para inferir haplotipos en pedigrís de tamaño moderado (menor a 40 individuos por familia), y la tabla 2, los principales algoritmos para inferir haplotipos en pedigrís complejos y de gran tamaño.

Tabla 1. Métodos principales de Inferencia de Haplotipos en pedigrís de tamaño moderado (<40).

Nombre del programa	Algoritmos usados	Asume no recombinaciones	Limitaciones	Configuración identificada	Referencias
SUPERLINK	Red bayesiana	No	Pocos marcadores, con pedigrís de hasta algunos centenares	Máxima verosimilitud	[24]
GENEHUNTER	Lander-Green	No	Gran número de marcadores	Máxima verosimilitud	[17]
Merlín	Lander-Green	No	Cientos de marcadores	Máxima verosimilitud	[19]
Allegro 2	Lander-Green	No	Cientos de marcadores	Máxima verosimilitud	[18]
ZAPLO	EM	Si	Pocos marcadores	Cero recombinaciones	[25]
HAPLORE	EM	Si	Pocos marcadores	Cero recombinaciones	[26]
PhyloPed	Filogenia	No	Miles de marcadores	Máxima verosimilitud	[27]

Tabla 2. Métodos principales de Inferencia de Haplotipos en pedigrís grandes y complejos.

Nombre del programa	Algoritmos usados	Asume no recombinaciones	Limitaciones	Configuración identificada	Referencias
SimWalk 2	Recocido simulado	No	Gran número de marcadores, lento	Máxima verosimilitud	[20]
PedPhase	EM, Programación Lineal Entera	No	Gran número de marcadores	Mínimo número de recombinaciones	[28]
reHCstar	Solucionador SAT	No	Cientos de marcadores	Mínimo número de recombinaciones	[29]
TDS 2	Ligadura de partición	No	Gran número de marcadores	Mínimo número de recombinaciones	[30]
SIMPLE	Imputación secuencial	No	Pequeño o moderado número de marcadores	Máxima verosimilitud	[31]
CeHap	Enumeración condicional	No	Gran número de marcadores, no puede procesar valores faltantes	Máxima verosimilitud	[32]

Tablas recuperadas y modificadas de [33].

2.4.1 Enfoques para la inferencia de haplotipos en pedigrís

2.4.1.1 Configuración de Haplotipos con Cero Recombinaciones (ZRHC)

Dado un grafo pedigrí G de genotipos válido, encontrar una solución H de G que no implique eventos de recombinación o decidir qué tal solución no existe.

2.4.1.2 Configuración de Haplotipos de Recombinación Mínima (MRHC)

Dado un grafo pedigrí de genotipos válido G, encontrar una solución H de G que implique un número mínimo de eventos de recombinación.

2.4.1.3 Configuración de haplotipos con k recombinaciones (k-MRHC)

Dado un grafo pedigrí de genotipos válido G, encontrar una solución H de G tal que el número total de recombinaciones sea mínimo y el número de recombinaciones en cada par padre-hijo sea como máximo k [34].

2.5 Inferencia en individuos no relacionados

En esta variante del problema HI, se debe reconstruir (o estimar) las secuencias de ADN de los dos haplotipos a partir de un cromosoma (o un segmento fijo del mismo) de un solo individuo específico, del cual no se tiene información de sus padres. Esta variante con frecuencia surge cuando se hace una secuenciación (relativamente rápida y barata) del genoma (o partes del mismo) de un humano individual.

En esa tecnología, las lecturas de secuenciación son cortas, con una alta tasa de error, pero existe un "genoma de referencia" para los seres humanos, y se utiliza en la reconstrucción.

La entrada a esta variante del problema HI es un conjunto de lecturas (secuencias cortas de ADN) de diferentes intervalos de los dos haplotipos (desconocidos). Las secuencias de haplotipos se denotan H_0 y H_1. Cada lectura individual es una secuencia que proviene sólo de uno de los haplotipos, pero, para dos lecturas cualquiera, no se sabe si las lecturas se originan a partir del mismo haplotipo, o de los dos haplotipos diferentes. Por lo tanto, no es inmediato cómo ensamblar las secuencias de ADN de las lecturas en dos secuencias largas, H_0 y H_1, para estimar mejor los verdaderos haplotipos H_0 y H_1.

El problema clave es cómo dividir el conjunto de lecturas en dos conjuntos, con la interpretación de que las lecturas de un conjunto proceden de un haplotipo y las lecturas del otro conjunto proceden del otro haplotipo. Esta división se utiliza para crear H_0 y H_1. Suponemos que conocemos la posición de cada lectura, en relación con la longitud del cromosoma (o segmento de la misma) que se secuencia. Por ejemplo, podríamos saber que el inicio de una lectura en particular se encuentra a 300 bases del extremo de 5' del cromosoma de interés. Esta suposición es realista para los seres humanos y otras especies donde hay un genoma de referencia terminado disponible. Es posible que las lecturas individuales no estén completamente de acuerdo con la referencia (especialmente porque solo hay una secuencia de referencia, pero hay dos haplotipos no idénticos), pero el acuerdo entre las lecturas y la secuencia de referencia es suficiente para determinar la ubicación correcta de la lectura [16].

2.6 Software existente para HI en tríos y dúos.

2.6.1 Beagle 5.1

Es un software para la inferencia de haplotipos e imputación de datos de genotipos faltantes, para la implementación de estudios de asociación. Trabaja con datos de genotipos desfasados de individuos no relacionados, tríos, y pares. Cuenta con una precisión probada del 99% para inferir datos reales de 3,002 individuos genotipificados para 490,032 marcadores, en 3.1 días de tiempo de computación.

Beagle utiliza un modelo de LD (Desequilibrio de Ligamiento) [43] empírico, que es una clase especial de grafo acíclico dirigido, que se adapta a la estructura local de los datos. En relación con otros métodos, funciona particularmente bien con tamaños de muestra grandes. Puede utilizarse para muestrear pares de haplotipos o para encontrar el par de haplotipos más probable para cada individuo condicionado a los genotipos observados [50].

2.6.2 GenHap

GenHap es un algoritmo basado en Algoritmos Genéticos para el ensamblaje de haplotipos, desarrollado para gestionar lecturas largas.

GenHap aborda la complejidad computacional del problema del haplotipo explotando los Algoritmos Genéticos para resolver el problema de la Corrección Mínima de Errores ponderada (wMEC), una variante del conocido problema MEC (Corrección Mínima de Errores). GenHap puede resolver eficientemente grandes instancias del problema wMEC, produciendo soluciones óptimas mediante un proceso de búsqueda global, sin ninguna hipótesis a priori sobre la

distribución de errores de secuenciación en las lecturas. Además, GenHap es capaz de procesar conjuntos de datos compuestos por lecturas largas y coberturas de hasta 60x en ordenadores personales [35].

2.6.3 ShapeIt 4

Su enfoque consiste en reunir pequeños conjuntos de haplotipos informativos sobre los que condicionar la estimación de los haplotipos. Dichos de haplotipos son una estructura de datos en la que dos haplotipos que comparten un prefijo largo (es decir, que coinciden) en una posición determinada se ordenan uno al lado del otro en esa posición. *SHAPEIT4* se aprovecha de ello manteniendo todas las estimaciones de los haplotipos para poder identificar las coincidencias largas entre los haplotipos en tiempo constante.

En la práctica, *SHAPEIT4* trabaja dentro de regiones genómicas superpuestas (de 2 Mb por defecto) y procede como sigue para actualizar la fase de un individuo en una región determinada (i) interroga las matrices cada ocho variantes para obtener los haplotipos *P* que comparten los prefijos más largos con las estimaciones actuales de haplotipos en esa posición, (ii) colapsa los haplotipos identificados en toda la región en una lista de *K* haplotipos distintos, y (iii) ejecuta el condicionamiento en los *K* haplotipos [36].

Capítulo III

3.1 Materiales y Métodos

Los algoritmos seleccionados para este estudio se ejecutaron en las siguientes plataformas: *PedPhase* (*BE, ILP*) en Windows 10, *ReHCstar* y *SimWalk* en la distribución Ubuntu 20.04 en Linux. Todos los programas se ejecutaron con los parámetros predeterminados por la herramienta.

Dentro del conjunto de algoritmos encontrados en el estado del arte para inferencia de haplotipos en pedigrís, se encuentran *Merlín*, y *Allegro*, sin embargo, el primero no pudo manejar la cantidad de individuos en la familia del pedigrí analizado (79 miembros), teniendo dicho algoritmo un límite de 27 individuos no fundadores. El algoritmo *Allegro* presentó un error de compatibilidad en el OS Windows 10, y al no tener existencia de una versión para sistemas basados en Unix, se prescindió de dicho programa. Por otra parte, los algoritmos *Haplore* y *Zaplo* no arrojaron resultados debido a que el pedigrí incluye recombinaciones en los datos, característica no permitida por dichos programas. De manera similar, PhyloPed no tuvo la capacidad de procesar el pedigrí, dado que cuenta con lazos de consanguinidad entre sus individuos. Finalmente, el algoritmo llamado *GeneHunter* no presentó soporte a los enlaces de acceso para su descarga, lo cual imposibilitó su utilización.

3.1.1 Simulación de datos de genotipos.

Debido a la necesidad de evaluar diversos métodos estadísticos en genómica, es común que se generen simulaciones de datos de variantes genéticas, con el fin de estudiar el comportamiento de dichos algoritmos ante diferentes características relevantes de los datos.

Existen programas para la generación de datos de genotipos, tales como *SLINK* [37], *ALLEGRO* [18], o paquetes para el programa R, como lo es *SIM1000G* [38]. Es posible generar datos de marcadores en equilibrio de ligamiento, con o sin rasgos fenotípicos, respetando el desequilibrio de Hardy-Weinberg [42], y estructuras configurables de pedigrís, sin embargo, algunos de estos programas presentan ciertas limitantes, como la generación de pedigrís de baja complejidad, y la imposibilidad de producir un gran número de marcadores. Debido a ello, se debe seleccionar el programa de simulación adecuado, que nos ayude a cumplir con las necesidades del análisis a realizar.

En este estudio, se utilizó el programa *SimPed* [39], el cual se explica en el apartado 3.1.2, para la simulación de un pedigrí con datos de genotipos, de 79 individuos, con 500 marcadores SNP, con la estructura que se presenta en la figura 6.

Se trabajó bajo los siguientes recursos computacionales:

Procesador: AMD Ryzen 5 3500U, 2.10 GHz, x64

RAM: 12.0 Gb.

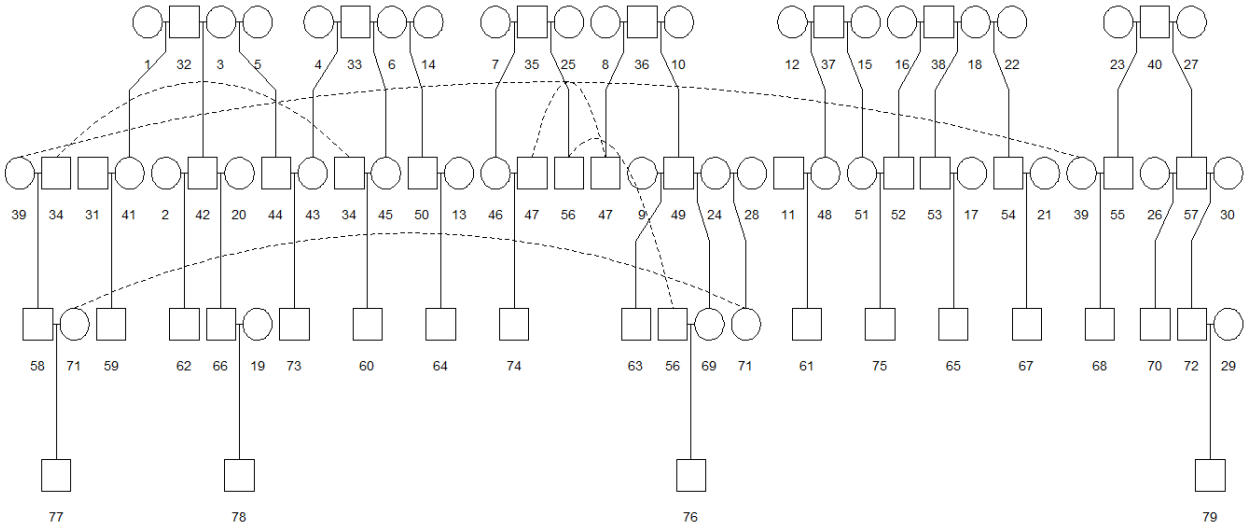


Figura 6. "Diagrama del pedigrí simulado, 79 miembros de los cuales 40 son fundadores y 39 no fundadores. Pedigrí dibujado con el paquete en R llamado Kinship2 [40]."

3.1.2 SimPed

Es un programa de simulación que genera datos de haplotipos y/o genotipos con o sin rasgos fenotípicos para pedigrís de virtualmente cualquier tamaño y complejidad. El programa puede generar datos para un largo número de marcadores loci bialélicos o multialélicos. (>20,000 loci). Los genotipos pueden ser generados bajo la asunción del equilibrio de Hardy-Weinberg [42], proveyendo las frecuencias alélicas. SimPed acepta datos de pedigrí en formato *PLINK* estándar. Los datos de haplotipos y/o genotipos pueden generarse tanto para los autosomas como para el cromosoma X.

El programa SimPed genera datos de haplotipos y/o genotipos para los pedigríes de la siguiente manera. Para los autosomas, a todos los fundadores del pedigrí se les asignan dos haplotipos y/o alelos condicionados en función de las frecuencias especificadas para todos los marcadores. Empezando por la parte superior de la estructura del pedigrí, a la primera descendencia del

fundador se le asigna aleatoriamente uno de los haplotipos. A continuación, se determina, basándose en el mapa genético, si se ha producido un evento de recombinación entre el primer y el segundo marcador. Si con probabilidad se ha producido un evento de recombinación, entonces en el segundo marcador locus se asigna a la descendencia el alelo del otro haplotipo del fundador. Si no se ha producido un evento de recombinación con probabilidad (1 - Θ), se asigna el alelo del mismo haplotipo del fundador en el segundo marcador locus. Este procedimiento se repite hasta que los alelos de todos los loci de los marcadores se hayan asignado de un fundador a su descendencia. El proceso se repite, esta vez asignando los alelos a la descendencia de su otro progenitor.

SimPed requiere de dos archivos de entrada para funcionar, un archivo de texto con extensión *.pre*, con la estructura de pedigrí a generar, y un archivo de texto en extensión *.dat*, con los parámetros de entrada.

Los parámetros dados en el archivo de entrada para la generación de los genotipos en el pedigrí fueron los siguientes: 0 columnas de estado de afección o rasgo cuantitativo, y 0 para generar datos autosómicos (Columna 3 del archivo de entrada *.dat*), 1 replica a generar (Columna 4), 500 marcadores (Columna 5), 2 para colocar las distancias genéticas (Columna 6) y en el siguiente renglón (Columna 7) se colocó 499 (# de marcadores - 1), y las respectivas distancias. La Columna 8 con el valor 2, 500, y 1 para producir genotipos, 500 de ellos, y repetir el patrón una vez, respectivamente. Finalmente, en la columna 9, se enlistó un 2 para cada marcador, (para indicar que serán 2 alelos por marcador), y su respectiva frecuencia alélica del primer marcador.

3.1.3 Formatos de archivos de entrada

3.1.3.1 Archivo de Pedigrí

El formato *PLINK* estándar de pedigrí que se utiliza ampliamente es el archivo PED, el cual es un archivo delimitado por espacios en blanco (espacio o tabulación): las seis primeras columnas son obligatorias:

- ID familiar
- ID del individuo
- ID Paterno
- ID Materno
- Sexo (1=hombre; 2=mujer; otro=desconocido)
- Fenotipo

Los IDs son alfanuméricos: la combinación del ID de la familia y del individuo debe identificar de forma única a una persona. Un archivo PED debe tener 1 y sólo 1 fenotipo en la sexta columna. El fenotipo puede ser un rasgo cuantitativo o una columna de estado de afección [42].

3.1.3.2 Archivo de Mapa

Cada línea del archivo MAP describe un único marcador y debe contener exactamente 4 columnas:

- Cromosoma (1-22, X, Y o 0 si no se conoce)
- rs# o ID del SNP
- Distancia genética (en Morgans)
- Posición en bases par (en unidades bp)

Se espera que las posiciones de pares de bases correspondan a números enteros positivos dentro del rango de los tamaños típicos de los cromosomas de la especie a analizar [41].

3.1.4 Equilibrio de Hardy-Weinberg

Nos dice que las frecuencias de los genotipos dentro de una población tienden a permanecer en equilibrio. Esta regla, aparentemente sencilla, establece que para un locus bialélico con frecuencias p y q respectivamente, las frecuencias genotípicas deben seguir la ecuación:

$$p^2 + 2pq + q^2 = 1$$

Donde p^2 es el alelo homocigoto dominante, $2pq$ es el alelo heterocigoto, y q^2 es el alelo recesivo.

Las alteraciones del Equilibrio de Hardy Weinberg se producen cuando la selección natural opera sobre un genotipo concreto dando una aptitud diferencial a cualquiera de ellos.

Los genetistas determinan los genotipos y, a continuación, comparan las frecuencias alélicas en pacientes no relacionados y en pacientes de control. Siempre que el alelo mutante esté en desequilibrio de ligamiento con la mutación causante, se observa un aumento estadísticamente significativo de su frecuencia alélica entre los casos. Sin embargo, independientemente de la frecuencia alélica y del locus analizado, cualquier marcador genético que mapee el cromosoma autosómico deberá seguir el equilibrio de Hardy Weinberg [42].

3.1.5 Desequilibrio por ligamiento

El desequilibrio por ligamiento (LD) se define como la distribución no aleatoria de alelos en diferentes loci, es decir, que existe cierta relación medible entre alelos, y tienden a heredarse

juntos, presentando cierta correlación. La evaluación cuantitativa de la LD en una población de interés es un procedimiento importante para llevar a cabo el mapeo fino de las variantes causales integradas en los loci de riesgo de enfermedad identificados por estudios de asociación de todo el genoma (GWAS).

Las medidas de LD más utilizadas son r^2 y D'; ambos valores cuantifican el LD entre variantes bialélicas (es decir, SNPs) para dos sitios, reflejando distribuciones no aleatorias de cuatro haplotipos que consisten en combinaciones por pares de los alelos. Específicamente, r^2 puede interpretarse como la medida de correlación de Pearson de las distribuciones de alelos y se sabe que es proporcional a los valores χ^2 de las estadísticas de asociación genotipo-fenotipo entre dos sitios [43].

3.2 Base de datos de genotipos "WIDDE: Web-Interfaced next generation Database dedicated to genetic Diversity Exploration"

Es una base de datos de nueva generación con interfaz web que sirve como herramienta para una amplia gama de especies y tipos de marcadores. Cuenta con datos de biodiversidad de ganado bovino, que incluye datos de genotipos para más de 750,000 SNPs disponibles en 129 poblaciones bovinas. Es capaz de realizar un filtrado opcional de los datos, y la exportación de los mismos en los formatos más populares, así como la exploración de la diversidad genética a través de un análisis de componentes principales. Los usuarios también pueden explorar sus propios datos junto con los datos de WIDDE y asignar las muestras a las poblaciones de la base de datos [44].

3.3 Control de calidad de los datos

Se tomó el cromosoma 6, del ensamblaje UMD3.1, de una población de 63 individuos de ganado Holstein, con 2,557 marcadores obtenidos con el chip Illumina BovineSNP50v1. De este cromosoma, se seleccionaron 500 marcadores SNP, presentes en una distancia genómica de 26 Mb. Dichos genotipos fueron obtenidos de la base de datos mencionada anteriormente, y el ganado a su vez fue genotipificado como lo explica Matukumalli [45].

Utilizando la herramienta en línea de la base de datos, como se muestra en la figura 7, se aplicó un filtro de cobertura del 95% para los marcadores de los individuos, esto es, se eliminaron los individuos que presentaron menos del 95% de sus marcadores. Se eliminaron los marcadores cuyo genotipo fue conocido por menos del 75% de los individuos. Se estableció un valor p del 0.001 como el criterio para eliminar los marcadores que no cumplieran con el equilibrio de Hardy-Weinberg. Se eliminaron finalmente los marcadores cuyas frecuencias alélicas fueron menores a 5%.

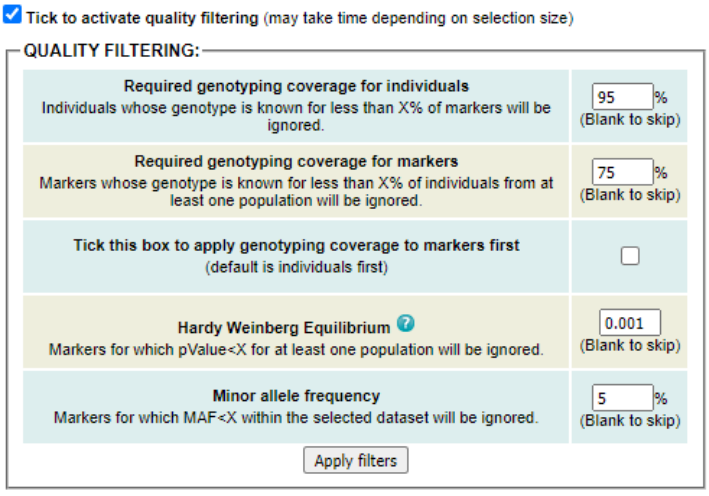


Figura 7. "Filtro de calidad aplicado a los datos de genotipos del cromosoma 6 del ganado Holstein UMD3.1 en la base de datos WIDDE."

3.3.1 Plink 1.9

PLINK es un conjunto de herramientas para los GWAS de código abierto en C/C++. Con *PLINK*, se pueden manipular y analizar rápidamente grandes conjuntos de datos que comprenden cientos de miles de marcadores genotipados para miles de individuos. Además de proporcionar herramientas para hacer que los pasos analíticos básicos sean computacionalmente eficientes, *PLINK* también admite algunos enfoques novedosos que aprovechan la cobertura del genoma completo. Cuenta con cinco funciones principales: gestión de datos, estadísticas de resumen, estratificación de la población, análisis de asociación y estimación de identidad por descendencia [46].

PLINK permite transformar archivos de datos de genotipos de un formato a otro, ya sea de VFC a *PED/MAP*, y viceversa, o a formatos ampliamente utilizados como BAM/FAM/BIM.

Se realizó el cálculo de las frecuencias alélicas para cada marcador del archivo de genotipos, con el siguiente comando en *PLINK*:

plink.exe --file cattle_2557variants_63individuals --freq --keep-allele-order

3.4 Algoritmos para la inferencia de haplotipos en Pedigrís

3.4.1 PedPhase ILP

El algoritmo *ILP* (*Programación Lineal Entera*) formula el problema MRHC con datos faltantes utilizando la técnica de programación lineal entera, emplea una estrategia de ramificación y limitación que utiliza una relación de orden parcial y algunas otras relaciones especiales entre las variables para decidir el orden de ramificación. La relación de orden parcial se descubre en el

preprocesamiento de las restricciones considerando propiedades únicas en la formulación. Se construye un gráfico dirigido basado en las variables y su relación de orden parcial. Al identificar y colapsar los componentes fuertemente conectados en el gráfico, podemos reducir en gran medida el tamaño de una instancia de *ILP*. Se introducen límites inferiores y superiores no triviales en el número óptimo de recombinantes que se introducen en cada nodo de bifurcación para podar eficazmente el árbol de búsqueda. Cuando existen múltiples soluciones, se selecciona la mejor configuración de haplotipos basándose en un enfoque de máxima probabilidad [28].

3.4.2 PedPhase Extensión de Bloques

El algoritmo de extensión de bloques intenta resolver la configuración de los haplotipos de todos los loci no ambiguos utilizando la ley de herencia mendeliana. A continuación, se utilizan técnicas como evitar las dobles recombinaciones dentro de una pequeña región, para resolver los loci que son adyacentes a los loci previamente resueltos, dando lugar a bloques de loci resueltos consecutivos. Posteriormente, el algoritmo utiliza el bloque más largo del pedigrí para resolver más loci, según el principio de mínima recombinación. El proceso se repite hasta que no se puede ampliar ningún bloque. Después el algoritmo rellena los bloques en cada miembro teniendo en cuenta la información del haplotipo en los otros miembros de la misma familia nuclear. La complejidad del algoritmo es $O(dmn)$, donde n es el tamaño del pedigrí, m el numero de loci, y d el mayor número de hijos en una familia [47].

3.4.3 SimWalk2

SimWalk2 es una aplicación informática de genética estadística para realizar análisis de haplotipos, ligamiento paramétrico, ligamiento no paramétrico (NPL), identidad por descendencia (IBD) y mistyping en pedigríes de cualquier tamaño. *SimWalk2* utiliza los algoritmos Markov Chain Monte Carlo (MCMC) y Recocido Simulado para realizar estos análisis. El algoritmo MCMC es capaz de analizar grandes pedigríes porque considera las configuraciones subyacentes en proporción a su probabilidad. Así, una configuración que es teóricamente posible pero muy poco probable (debido al gran número de recombinaciones que requeriría esa configuración) a menudo no se tendrá en cuenta.

El análisis de haplotipos estima el conjunto más probable de haplotipos maternos y paternos completamente tipificados de los marcadores en cada individuo del pedigrí. Se destacan los eventos de recombinación dentro de los haplotipos. Se proporcionan varias medidas para indicar la probabilidad de estas recombinaciones que pueden exponer errores de genotipificación.

Implementa un gráfico de descenso que especifica las trayectorias del flujo de genes, pero no los fundadores concretos que viajan por las trayectorias. Estima las puntuaciones de localización, y calcula las estadísticas de agrupación de genes para un análisis de ligamiento [20].

3.4.4 ReHCStar

ReHCStar, propone el método Minimum-Recombinant Haplotype Configuration with bounded Errors (MRHCE), que amplía la formulación original de Configuración de Haplotipos Mínimos Recombinantes (MRHC) incorporando las dos características más comunes de los datos reales: los errores y los genotipos faltantes (incluyendo los individuos no genotipificados). Se basa en una

reducción del conocido problema de Satisfacibilidad (SAT) y explota los recientes avances en la literatura de la programación de restricciones.

La descripción del enfoque se compone de dos partes: la primera parte, describe una reducción polinómica del problema $(r - e)$-HC al problema de satisfacibilidad (SAT), en la segunda parte, el algoritmo de MRHCE calcula el número mínimo de recombinaciones necesarias para resolver la instancia (permitiendo como máximo e errores de genotipado) imitando el comportamiento de una búsqueda binaria e invocando SOLVE_REHC, un solucionador SAT [29].

Capítulo IV

4.1 Resultados y Discusión

4.1.1 Criterios de evaluación para los algoritmos

Para analizar el desempeño de los métodos para inferir haplotipos, se toman en cuenta principalmente el tiempo de ejecución en función del tamaño de los datos de entrada, debido a que, basándonos en este dato, podemos decidir si el algoritmo se ajusta o no a nuestras necesidades en cuanto al tiempo de espera de los resultados. No menos importante es la precisión en la recuperación de la configuración de los haplotipos en comparación con los haplotipos reales, siendo este criterio de suma importancia para valorar la sensibilidad que se requiera alcanzar en los estudios posteriores. Los errores de conmutación (o errores de switch, SE) nos dicen el número de switches o cambios entre sitios heterocigotos vecinos necesarios en los haplotipos inferidos para recuperar la configuración de haplotipos original, y se formula mediante la Tasa de Errores de Conmutación o Switch Error Rate (SER). Formalmente se expresa de la siguiente manera:

$$SER = \frac{\sum_{i=1}^{n} t_i}{\sum_{i=1}^{n} (e_i - 1)}$$

donde t_i es el número de switches necesarios en el genotipo i para recuperar la solución real, y e_i es el número total de sitios heterocigotos en el genotipo i [48]. El cálculo se aplica a cada reconstrucción de haplotipos individual. El número de recombinaciones, por otra parte, se evalúa comúnmente dependiendo de la naturaleza de los algoritmos, siendo éstos últimos, por ejemplo, evaluados sólo en los algoritmos que acepten posibles recombinaciones para la generación de los

haplotipos presentes en las soluciones. Como se estudió, estos algoritmos que permitan recombinaciones, no son algo común.

4.1.1.1 Tiempo de ejecución

ReHCstar y *PedPhase* (*Extensión de Bloques* y *ILP*), miden por cuenta propia el tiempo de ejecución de sus programas, por tanto, en SimWalk, se calculó el tiempo de ejecución mediante el uso de la terminal en Ubuntu.

4.1.1.2 Precisión

Se mide contando el porcentaje de cambios ocurridos en los haplotipos (materno y paterno) con respecto a los haplotipos reales. La proporción total de genotipos, menos este porcentaje, nos otorga la proporción de genotipos cuya fase fue inferida correctamente.

4.1.1.3 Errores de Switch

Un error de switch ocurre cuando un sitio heterocigoto es inferido de manera incorrecta con respecto al heterocigoto anterior. En la figura 8, los cuadrados de color naranja y azul representan 8 alelos heredados por la madre y el padre, respectivamente. Se aprecia un switch o cambio en la fase correcta en la posición 4 de los haplotipos centrales, representando un switch simple, y en los haplotipos de la derecha se aprecian dos switches, en la posición 4 y en la 5, los switches dobles [49]. En este estudio se tomaron en cuenta ambos tipos de switches.

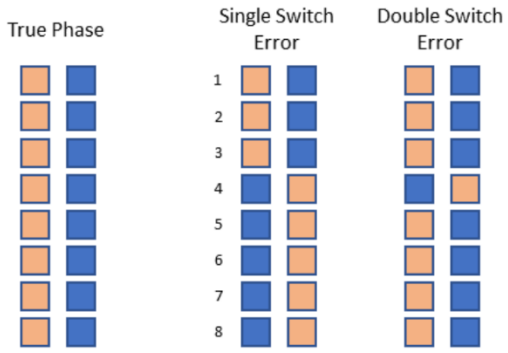


Figura 8. "Representación esquemática de los errores de switch. Cada columna representa una estimación de haplotipos de ocho genotipos heterocigotos."

4.1.1.4 Recombinaciones

Si tomamos en cuenta el criterio de mínima recombinación en haplotipos, el número de éstas se vuelve especialmente importante, ya que, a menor número de recombinaciones presentes en la configuración de haplotipos, mayor será la probabilidad de que ésta sea la solución con mayor realismo biológico.

Se ejecutaron los 4 algoritmos: *PedPhase* (*Extensión de Bloques*), *PedPhase* (*ILP*), *ReHCstar* y *SimWalk* en un pedigrí conformado por 79 individuos de los cuales 40 son fundadores y 39 son no fundadores, a su vez el conjunto de datos de 500 marcadores SNP se dividió en conjuntos de 5, 10, 20 y 50 SNPs para evaluar el desempeño de los algoritmos en función de la longitud de los datos, capturándose la precisión, como se muestra la siguiente tabla:

Tabla 3. Resultados de la precisión en los cuatro algoritmos evaluados.

# de SNPs	***Precisión (%)***			
	BE	***ILP***	***ReHCstar***	***SimWalk***
5	*82.27*	*82.27*	*75.69*	*75.94*
10	*79.87*	*79.62*	*75.44*	*68.22*
20	*77.46*	*77.15*	*74.68*	*67.59*
50	*76.48*	*76.48*	*76.0*	*65.31*

Para la obtención de los parámetros de precisión, número de errores de switch y precisión de switch, se utilizó el lenguaje de programación Python (https://www.python.org/), con el uso de las librerías Pandas (https://pandas.pydata.org/) y Numpy (https://numpy.org/), para el análisis de datos. Se calculó la precisión obteniendo el porcentaje de sitios inferidos correctamente con respecto a los haplotipos reales. Para hacer posible el cálculo de la precisión y los errores de switch, se asumió que los genotipos generados con SimPed eran los haplotipos reales.

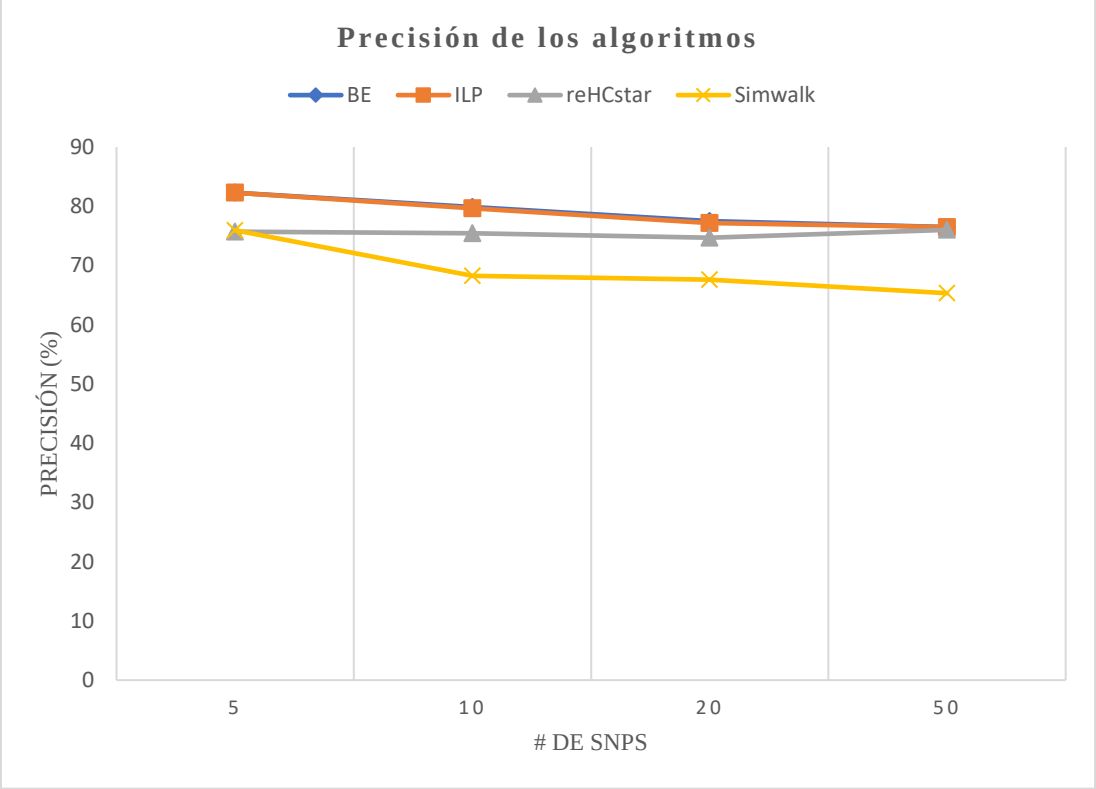


Figura 9. "Precisión respecto al número de SNPs, de los cuatro algoritmos evaluados, BE, ILP, ReHCstar y SimWalk."

Se observa que la precisión general de los algoritmos en este escenario, oscila entre 65.31% y 82.27%. A lo largo de todas las longitudes de los genotipos, los algoritmos más precisos fueron *BE* y *ILP*, y el menos preciso fue *SimWalk*. Se destaca la estabilidad en la precisión de *ReHCstar*.

El tiempo de ejecución se midió en segundos (ver tabla 4) y se capturó de manera automática en los algoritmos *BE*, *ILP* y *ReHCstar*, en cuanto a *SimWalk*, se calculó mediante la terminal en Ubuntu con el comando *time*, al momento de ejecutar el programa.

Tabla 4. Resultados del tiempo de ejecución, en segundos, en los cuatro algoritmos evaluados.

# de SNPs	***Tiempo de ejecución (segundos)***			
	BE	***ILP***	***ReHCstar***	***SimWalk***
5	*0.031*	*0.312*	*~ 0*	*258.9*
10	*0.062*	*0.547*	*~ 0*	*586.6*
20	*0.031*	*10.9*	*~ 0*	*1338.6*
50	*0.062*	*91586.3*	*7*	*12066.6*

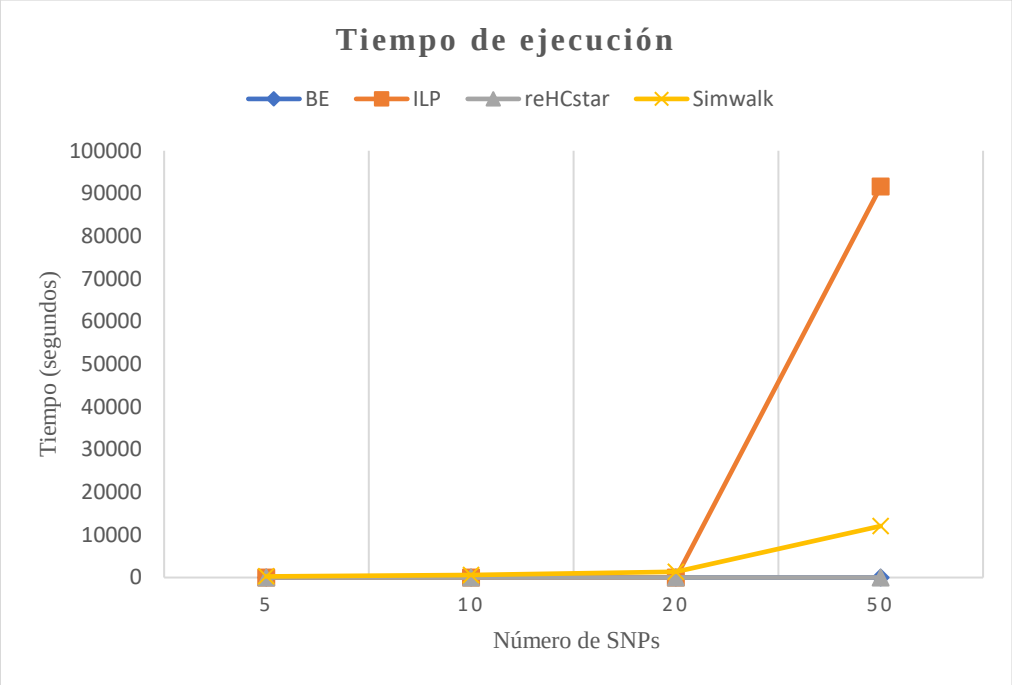


Figura 10. "Tiempo de ejecución respecto al número de SNPs de los cuatro algoritmos evaluados."

Se aprecia que el algoritmo más rápido, independientemente de la longitud de los datos es *Extensión de Bloques* (*BE*), siendo éste un algoritmo que trabaja en un tiempo casi lineal a lo largo de grandes cantidades de SNPs, sin embargo, está sujeto a un límite de 150 SNPs aproximadamente, en las condiciones computacionales que se utilizaron para este estudio. De

manera similar el algoritmo *ReHCstar* resulta bastante rápido, aunque se aprecia un ligero incremento del tiempo de ejecución a partir de 50 SNPs, alcanzando un tiempo de ~45 minutos en un conjunto de 200 SNPs con recombinaciones mínimas (dicha ejecución no se tomó en cuenta para esta evaluación). *ILP* trabaja en un tiempo casi lineal, sin embargo, al llegar a 50 SNPs, el tiempo de ejecución se vuelve exponencial, alcanzando las ~25 horas. Finalmente, *SimWalk* es el algoritmo más lento, demorando ~4 minutos en procesar 5 SNPs, y ~3.3 horas en procesar 50 SNPs.

Tabla 5. Número de recombinaciones en las soluciones de cada algoritmo.

# de SNPs	***Número de recombinaciones***			
	BE	***ILP***	***ReHCstar***	***SimWalk***
5	*5*	*4*	*4*	*106*
10	*13*	*8*	*8*	*238*
20	*44*	*24*	*22*	*477*
50	*166*	*125*	*80*	*1158*

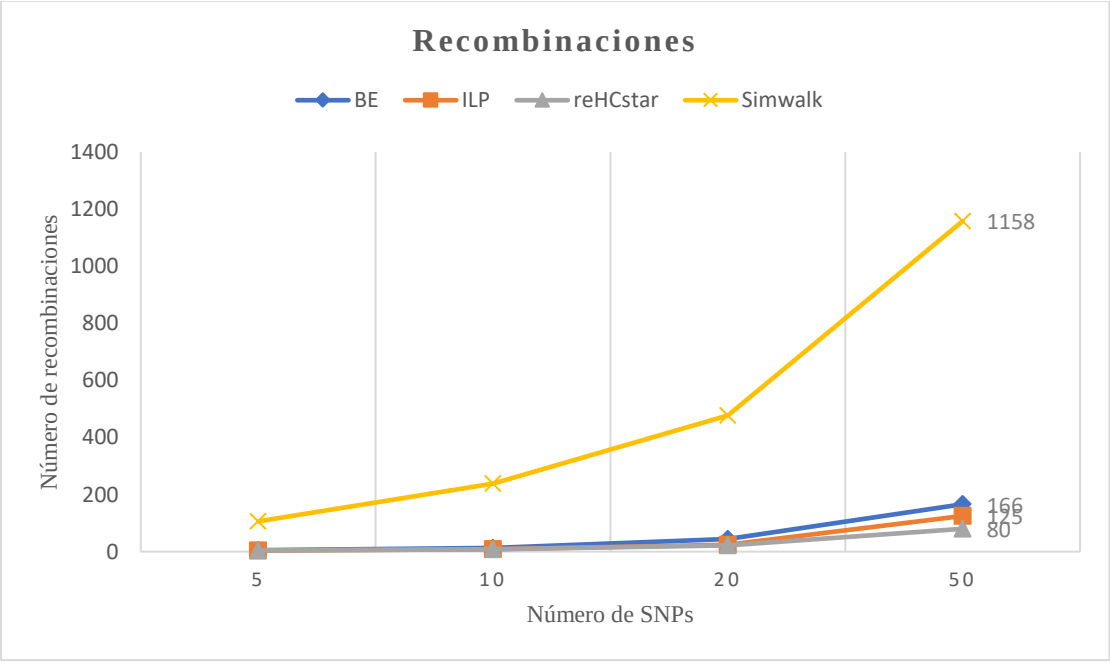


Figura 11. "Número de recombinaciones respecto al número de SNPs de los cuatro algoritmos evaluados."

Si tomamos en cuenta el principio de mínima recombinación, podemos observar en la tabla 5 y figura 11 que el *ReHCstar* cuenta con el menor número de recombinaciones a lo largo de los conjuntos de SNPs, y como se ha analizado en estudios anteriores [28], *SimWalk* tiende a sobrestimar el número de recombinaciones. *BE* y *ILP* trabajan de manera similar, calculando un número parecido de recombinaciones, 166 y 125, respectivamente para 50 SNPs.

Tabla 6. Número de Switch Errors (SE).

# de SNPs	***SE***			
	BE	***ILP***	***ReHCstar***	***SimWalk***
5	*13*	*12*	*11*	*11*
10	*68*	*64*	*63*	*74*
20	*258*	*260*	*254*	*221*
50	*868*	*846*	*846*	*596*

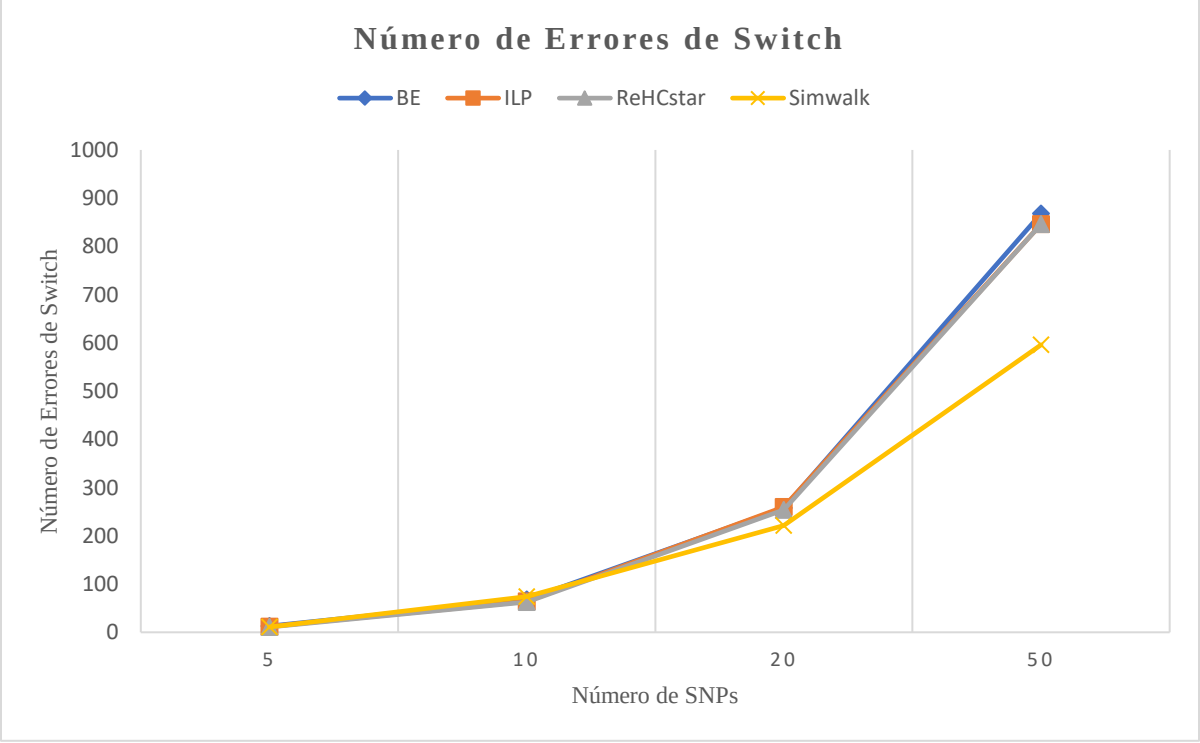


Figura 12. "Switch Errors en las soluciones de cada algoritmo, a lo largo del conjunto de SNPs."

Los errores de switch son un parámetro utilizado generalmente para evaluar la precisión de los algoritmos en cuanto a la inferencia de haplotipos. Un número menor de switches indica una mayor

precisión. Como se aprecia en la figura 12 y la tabla 6, *SimWalk* obtuvo un número menor de switches 596, en comparación con los demás algoritmos en 50 marcadores, *BE*: 868, ILP: 846 y ReHCstar: 846.

Tabla 7. Cálculo del Switch Error Rate (SER).

# de SNPs	***SER***			
	BE	***ILP***	***ReHCstar***	***SimWalk***
5	*0.094*	*0.086*	*0.082*	*0.077*
10	*0.199*	*0.192*	*0.189*	*0.224*
20	*0.368*	*0.367*	*0.357*	*0.315*
50	*0.473*	*0.462*	*0.460*	*0.323*

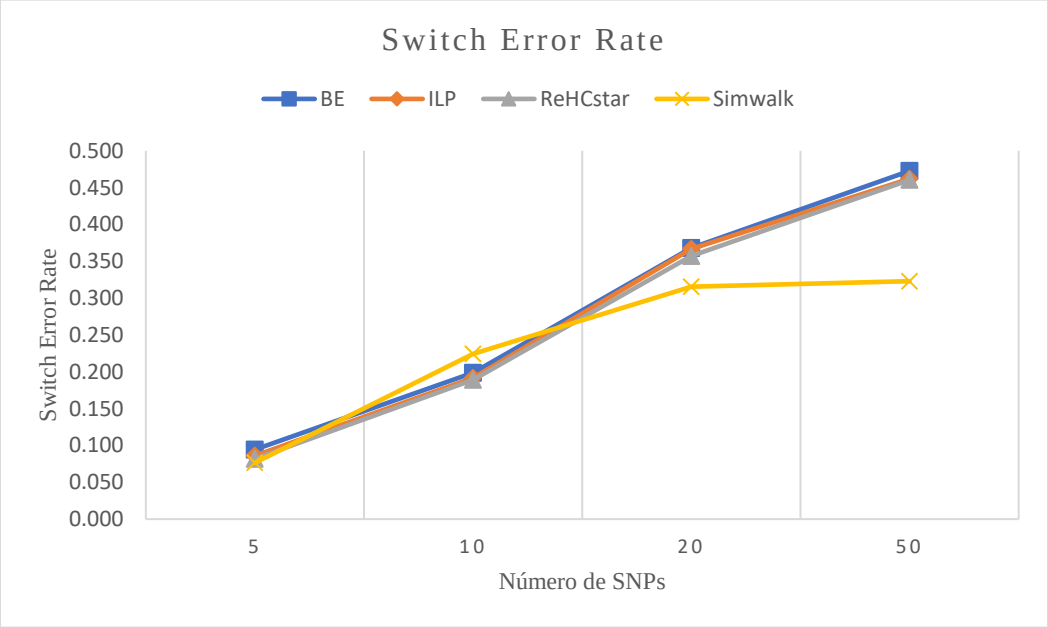


Figura 13. "Switch Error Rate en las soluciones de cada algoritmo, a lo largo del conjunto de SNPs."

La proporción de switches, o switch error rate no depende de la longitud de los datos, por tanto, es un buen indicador de la precisión de los algoritmos, independientemente de la cantidad de datos a procesar, ya que sólo toma en cuenta los sitios heterocigotos en comparación con la correcta inferencia respecto a los haplotipos reales. En la figura 13 se aprecia que *SimWalk* cuenta con una proporción menor de errores de switch, seguido de *ReHCstar*, *ILP* y finalmente *BE*.

Capítulo V

5.1 Conclusiones

En este trabajo de tesis se estudiaron y aplicaron métodos matemáticos para la estimación de haplotipos en pedigrís, partiendo de genotipos tipo SNP. Se ejecutaron ejercicios prácticos para un pedigrí con 79 individuos, utilizando 5, 10, 20 y 50 marcadores SNP. Se reportaron resultados de cuatro algoritmos: *Programación Lineal Entera*, *Extensión de bloques*, *ReHCstar* y *SimWalk*. Los criterios de evaluación para la inferencia de haplotipos fueron: el tiempo de ejecución del algoritmo, la precisión comparando los haplotipos inferidos con los haplotipos reales, la tasa de errores de conmutación, y el número de recombinaciones impuestas a los datos para realizar la reconstrucción de los haplotipos inferidos. Las ejecuciones fueron realizadas en una computadora personal con un procesador de 64 bits AMD Ryzen 5 3500U, con una velocidad de procesado de 2.10 GHz.

Los resultados muestran que los algoritmos que tienen la mejor precisión para los datos utilizados en este trabajo fueron *Extensión de Bloques* y *Programación Lineal Entera*, cuyas precisiones estuvieron en un rango de 76.48% a 82.27%. En cuanto a tiempo de ejecución, el algoritmo que presentó mejores resultados fue *ReHCstar*, cuy rango anduvo entre ~0 y 7 segundos (aunque en general todos los algoritmos fueron rápidos, y presentaron tiempos entre segundos y minutos). En cuanto al número de recombinaciones impuestas a los haplotipos inferidos, el algoritmo con mejores resultados fue *ReHCstar*, cuyo número de recombinaciones fue entre 4 y 80. Y en cuanto al número de errores de conmutación (Switch errors), el algoritmo con mejores

resultados fue también *ReHCstar*, cuyos valores fueron entre 41 y 926, que al estimar la tasa de errores resultó entre 0.337 y 0.504.

Debido a la existencia de pedigrís complejos en la industria ganadera, y dado el crecimiento acelerado en los últimos años de las tecnologías de genotipificación, es necesario el desarrollo y evaluación de algoritmos para inferir haplotipos con una alta precisión, y que trabajen con un tiempo de ejecución razonable, dado que los pedigrís de ganado bovino, en promedio cuentan con un alto número de individuos, desde unas cuantas decenas hasta centenares de ellos, por lo que; para lograr resultados significativos en posteriores estudios de asociación gen-enfermedad, entre otros análisis genéticos, es necesario el uso de un alto número de marcadores. Finalmente, en este estudio, el algoritmo que más cumplió con las exigencias para dicho fin fue *ReHCstar*, presentando un tiempo de ejecución corto en comparación con los algoritmos analizados, con la capacidad de manejar un alto número de marcadores SNPs, en el rango de las decenas de miles, minimizando el número de recombinaciones en sus soluciones (4, 8, 22 y 80 recombinaciones, para 5, 10, 20 y 50 marcadores SNPs), siendo el algoritmo que obtuvo una menor cantidad de las mismas en este estudio.

5.2 Trabajo a futuro

Tomando en cuenta el principio biológico del modelo de la inferencia de haplotipos de mínimas recombinaciones, *ReHCstar* podría llegar a ser un algoritmo factible para su aplicación en futuros análisis para la mejora genética de ganado en la región. En este caso se recomendará su utilización para la inferencia de haplotipos en datos de genotipos del proyecto de mejora genética de CONARGEN, que consta de decenas de miles de genotipos con información de pedigrís.

Referencias

[1] **Panorama General Del Proyecto Del Genoma Humano.** (s/f). *Genome.Gov*. Recuperado el 24 de febrero de 2021, de https://www.genome.gov/panorama-general-del-proyecto-del-genoma-humano.

[2] Pompanon, F., & Bonin, A. (Eds.). (2012). ***Data Production and Analysis in Population Genomics:*** *Methods and Protocols* (Vol. 888). Humana Press.

[3] Ramírez-Bello, J., Vargas-Alarcón, G., & Tovilla-Zárate, C. (s/f). **Polimorfismos de un solo nucleótido (SNP): Implicaciones funcionales de los SNP reguladores (rSNP) y de los SNP-ARN estructurales (srSNP) en enfermedades complejas.** Gaceta Médica de México., 9.

[4] Koopaee, H. K., & Koshkoiyeh, A. E. (2014). **SNPs genotyping technologies and their applications in farm animals breeding programs:** *Review*. *Brazilian Archives of Biology and Technology*, *57*(1), 87–95.

[5] Illumina (2021, marzo 24). **"Genotyping methods and solutions"**. https://www.illumina.com/techniques/popular-applications/genotyping/snp-snv-genotyping.html

[6] Khatkar, M. S., Zenger, K. R., Hobbs, M., Hawken, R. J., Cavanagh, J. A. L., Barris, W., McClintock, A. E., McClintock, S., Thomson, P. C., Tier, B., Nicholas, F. W., & Raadsma, H. W. (2007). **A primary assembly of a bovine haplotype block map based on a 15,036-single-nucleotide polymorphism panel genotyped in Holstein-Friesian cattle.** *Genetics*, *176*(2), 763–772.

[7] (Methods in Molecular Biology 1666) Robert C. Elston (eds.)—**Statistical Human Genetics:** *Methods and Protocols-Humana Press* (2017).pdf. (s/f).

[8] Browning, S. R., & Browning, B. L. (2011). **Haplotype phasing: Existing methods and new developments.** *Nature Reviews Genetics*, *12*(10), 703–714.

[9] Guttman, B., Griffiths, A., Suzuki, D., & Cullis, T. (2004). *Genetics A beginner's guide*.

[10] Istrail, S., Waterman, M. S., & Clark, A. G. (Eds.). (2004). **Computational methods for SNPs and Haplotype inference:** *DIMACS/RECOMB satellite workshop, Piscataway, NJ, USA, November 2002 revised papers / Sorin Istrail, Michael Waterman, Andrew Clark, (eds.)*. Springer-Verlag.

[11] Crawford, D. C., & Nickerson, D. A. (2005). **Definition and clinical importance of haplotypes.** *Annual Review of Medicine*, *56*, 303–320.

[12] Aguiar, D. (2014). **Genome-wide algorithms for haplotype assembly, haplotype phasing, and IBD inference.** 159.

[13] Frayling, T. (2014). **Genome-wide association studies: The good, the bad and the ugly.** *Clinical Medicine*, *14*(4), 428–431.

[14] Zhao, Y., Xu, Y., Zhang, Q., & Chen, G. (2007). **An overview of the haplotype problems and algorithms.** *Frontiers of Computer Science in China*, *1*(3), 272–282.

[15] Graça, A., Lynce, I., Marques-Silva, J., & Oliveira, A. L. (2010). **Haplotype inference by pure parsimony: A survey.** *Journal of Computational Biology*, *17*(8), 969–992.

[16] Gusfield D. (2019). **Haplotyping as perfect phylogeny: conceptual framework and efficient solutions**. In*: Proceedings of The Sixth Annual International Conference on Computacion.*

[17] Kruglyak, L., Daly, M. J., Reeve-Daly, M. P., & Lander, E. S. (1996). **Parametric and nonparametric linkage analysis:** A unified multipoint approach. *American Journal of Human Genetics*, *58*(6), 1347–1363.

[18] Gudbjartsson, D. F., Jonasson, K., Frigge, M. L., & Kong, A. (2000). **Allegro, a new computer program for multipoint analysis.** *Nature America Inc.*, *25*(may), 12–13.

[19] Abecasis, G. R., Cherny, S. S., Cookson, W. O., & Cardon, L. R. (2002). **Merlin — Rapid analysis of dense genetic maps using sparse gene flow trees.** *Nature Genetics*, *30*(1), 97–101.

[20] Sobel, E., & Lange, K. (1996). **Descent graphs in pedigrí analysis: Applications to haplotyping, location scores, and marker-sharing statistics.** *American Journal of Human Genetics*, *58*(6), 1323–1337.

[21] Mitchell, M. (2009). **An Introduction to Genetic Algorithms**. *MIT Press*.

[22] Angulo, R. V. (2009). **Improved computational methods for haplotype inference in large and complex pedigrís.** 1–23.

[23] Soeiro A (2011). **Satisfiability-based Algorithms for Haplotype Inference**. *Universidad Tecnica de Lisboa, Instituto Superior Técnico.*

[24] Fishelson, M., & Geiger, D. (2002). **Exact genetic linkage computations for general pedigrís.** *Bioinformatics*, *18*(SUPPL. 1).

[25] Rannala, B., & Slatkin, M. (2000). **Zero-recombinant haplotyping: Applications to fine mapping using SNPs**. *Genetic Epidemiology*, *19*(SUPPL. 1).

[26] Zhang, K., Sun, F., & Zhao, H. (2005). **HAPLORE: A program for haplotype reconstruction in general pedigrís without recombination.** *Bioinformatics*, *21*(1), 90–103.

[27] Kirkpatrick, B., Halperin, E., & Karp, R. M. (2010). **Haplotype inference in complex pedigrís.** *Journal of Computational Biology*, *17*(3), 269–280.

[28] Li, J., & Jiang, T. (2005**). Computing the minimum recombinant haplotype configuration from incomplete genotype data on a pedigrí by integer linear programming.** *Journal of Computational Biology*, *12*(6), 719–739.

[29] Pirola, Y., Vedova, G. Della, Biffani, S., Stella, A., & Bonizzoni, P. (2012). **A fast and practical approach to genotype phasing and imputation on a pedigrí with erroneous and incomplete information**. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *9*(6), 1582–1594.

[30] Iliadis, A., Anastassiou, D., & Wang, X. (2012). **A Unified Framework for Haplotype Inference in Nuclear Families.** *Annals of Human Genetics*, *76*(4), 312–325.

[31] Lin, S., Skrivanek, Z., & Irwin, M. (2003). **Haplotyping Using SIMPLE: Caution on Ignoring Interference.** *Genetic Epidemiology*, *25*(4), 384–387.

[32] Gao, G., & Hoeschele, I. (2008). **A rapid conditional enumeration haplotyping method in pedigrís.** *Genetics Selection Evolution*, *40*(1), 25–36.

[33] Gao, G., Allison, D. B., & Hoeschele, I. (2009**). Haplotyping methods for pedigrís.** *Human Heredity*, *67*(4), 248–266.

[34] Zhang, X.-S., Wang, R.-S., Wu, L.-Y., & Chen, L. (2008). **Models and Algorithms for Haplotyping Problem.** *Current Bioinformatics*, *1*(1), 105–114.

[35] Tangherloni, A., Spolaor, S., Rundo, L., Nobile, M. S., Cazzaniga, P., Mauri, G., Liò, P., Merelli, I., & Besozzi, D. (2019). **GenHap: A novel computational method based on genetic algorithms for haplotype assembly**. *BMC Bioinformatics*, *20*(S4), 172.

[36] Delaneau, O., Zagury, J.-F., Robinson, M. R., Marchini, J. L., & Dermitzakis, E. T. (2019). **Accurate, scalable and integrative haplotype estimation.** *Nature Communications*, *10*(1), 5436.

[37] Schäffer, A. A., Lemire, M., Ott, J., Lathrop, G. M., & Weeks, D. E. (2011). **Coordinated**

conditional simulation with SLINK and SUP of many markers linked or associated to a trait in large pedigrís. *Human Heredity*, *71*(2), 126–134.

[38] Dimitromanolakis, A., Xu, J., Krol, A., & Briollais, L. (2019). **sim1000G: A user-friendly genetic variant simulator in R for unrelated individuals and family-based designs.** *BMC Bioinformatics*, *20*(1), 1–9.

[39] Leal, S. M., Yan, K., & Müller-Myhsok, B. (2005). **SimPed: A simulation program to generate haplotype and genotype data for pedigrí structures.** *Human Heredity*, *60*(2), 119–122.

[40] Sinnwell, J. P., Therneau, T. M., & Schaid, D. J. (2014). **The kinship2 R package for pedigrí data.** *Human heredity*, *78*(2), 91–93.

[41] Purcell, S. (2010). **PLINK (1.07).** Documentation. http://zzz.bwh.harvard.edu/plink/dist/plink-doc-1.07.pdf. Accessed 28 Agosto. *Book*, 1–293.

[42] Royo, J. L. (2021). **Hardy Weinberg equilibrium disturbances in case-control studies lead to non-conclusive results.** *Cell Journal*, *22*(4), 572–574.

[43] Okada, Y. (2018**). eLD: entropy-based linkage disequilibrium index between multiallelic sites.** *Human Genome Variation*, *5*(1), 4–6.

[44] Sempéré, G., Moazami-Goudarzi, K., Eggen, A., Laloë, D., Gautier, M., & Flori, L. (2015). **WIDDE: A Web-Interfaced next generation database for genetic diversity exploration, with a first application in cattle.** *BMC Genomics*, *16*(1), 1–8.

[45] Matukumalli, L. K., Lawley, C. T., Schnabel, R. D., Taylor, J. F., Allan, M. F., Heaton, M. P., O'Connell, J., Moore, S. S., Smith, T. P. L., Sonstegard, T. S., & Van Tassell, C. P. (2009). **Development and Characterization of a High Density SNP Genotyping Assay for Cattle.** *PLoS ONE*, *4*(4).

[46] Purcell, Shaun, Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., De Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). **PLINK: A tool set for whole-genome association and population-based linkage analyses.** *American Journal of Human Genetics*, *81*(3), 559–575.

[47] Li, J. (2004). **PedPhase Version 2.0 User Manual.** *Bioinformatics*.

[48] Yang, J., Xu, Y., Yao, X., & Chen, G. (2013). **FNphasing: a novel fast heuristic algorithm for haplotype phasing based on flow network model.** *IEEE/ACM Transactions on*

Computational Biology and Bioinformatics / IEEE, ACM, *10*(2), 372–382.

[49] Browning, B. L., & Browning, S. R. (2022). **Genotype error biases trio-based estimates of haplotype phase accuracy.**

[50] Browning, S. R., & Browning, B. L. (2007). **Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering.** *The American Journal of Human Genetics*, *81*(5), 1084–1097.

Anexos

6.1 Estructura codificada del archivo pedigrí

```
1 1 0 0 2 1
1 2 0 0 2 1
1 3 0 0 2 1
1 4 0 0 2 1
1 5 0 0 2 1
1 6 0 0 2 1
1 7 0 0 2 1
1 8 0 0 2 1
1 9 0 0 2 1
1 10 0 0 2 1
1 11 0 0 1 1
1 12 0 0 2 1
1 13 0 0 2 1
1 14 0 0 2 1
1 15 0 0 2 1
1 16 0 0 2 1
1 17 0 0 2 1
1 18 0 0 2 1
1 19 0 0 2 1
1 20 0 0 2 1
1 21 0 0 2 1
1 22 0 0 2 1
1 23 0 0 2 1
1 24 0 0 2 1
1 25 0 0 2 1
1 26 0 0 2 1
1 27 0 0 2 1
1 28 0 0 2 1
1 29 0 0 2 1
1 30 0 0 2 1
1 31 0 0 1 1
1 32 0 0 1 1
1 33 0 0 1 1
1 34 0 0 1 1
1 35 0 0 1 1
1 36 0 0 1 1
1 37 0 0 1 1
1 38 0 0 1 1
1 39 0 0 2 1
1 40 0 0 1 1
1 41 32 1 2 1
1 42 32 3 1 1
1 43 33 4 2 1
1 44 32 5 1 1
1 45 33 6 2 1
1 46 35 7 2 1
1 47 36 8 1 1
1 48 37 12 2 1
1 49 36 10 1 1
1 50 33 14 1 1
1 51 37 15 2 1
1 52 38 16 1 1
1 53 38 18 1 1
1 54 38 22 1 1
1 55 40 23 1 1
1 56 35 25 1 1
1 57 40 27 1 1
1 58 34 39 1 1
1 59 31 41 1 1
1 60 34 45 1 1
1 61 11 48 1 1
1 62 42 2 1 1
1 63 49 9 1 1
1 64 50 13 1 1
1 65 53 17 1 1
1 66 42 20 1 1
```

```
1 67 54 21 1 1
1 68 55 39 1 1
1 69 49 24 2 1
1 70 57 26 1 1
1 71 49 28 2 1
1 72 57 30 1 1
1 73 44 43 1 1
1 74 47 46 1 1
1 75 52 51 1 1
1 76 56 69 1 1
1 77 58 71 1 1
1 78 66 19 1 1
1 79 72 29 1 1
```

6.2 Distancia de mapa genético utilizado en la simulación de los 500 genotipos, siendo 499 distancias. Valores en cM.

0.055578, 0.028226, 0.034191, 0.03737, 0.055047, 0.038234, 0.033982, 0.046286, 0.024486, 0.039255, 0.038875, 0.022947, 0.035588, 0.011986, 0.032966, 0.090597, 0.051049, 0.030949, 0.036111, 0.038264, 0.028377, 0.031718, 0.034454, 0.040373, 0.02757, 0.038653, 0.025077, 0.021488, 0.026159, 0.074857, 0.020389, 0.027695, 0.066686, 0.026557, 0.029325, 0.093963, 0.054979, 0.037331, 0.135927, 0.124594, 0.033029, 0.051327, 0.060839, 0.035093, 0.039992, 0.051068, 0.04051, 0.110582, 0.033288, 0.044932, 0.045873, 0.058069, 0.050634, 0.034166, 0.026382, 0.094107, 0.025877, 0.069036, 0.026503, 0.058637, 0.112413, 0.030172, 0.031658, 0.025424, 0.02184, 0.043669, 0.029912, 0.021897, 0.026649, 0.048186, 0.035606, 0.035966, 0.044756, 0.035999, 0.023767, 0.040092, 0.022465, 0.039862, 0.039156, 0.024669, 0.069334, 0.026901, 0.044654, 0.034651, 0.038505, 0.019517, 0.052979, 0.057679, 0.020212, 0.031218, 0.047059, 0.032979, 0.0232, 0.026166, 0.027681, 0.073612, 0.022727, 0.020446, 0.04319, 0.020244, 0.026735, 0.029388, 0.070973, 0.038361, 0.035054, 0.03631, 0.035475, 0.037219, 0.027843, 0.056994, 0.033683, 0.029035, 0.013555, 0.011906, 0.025054, 0.02303, 0.083902, 0.001858, 0.017541, 0.050447, 0.055342, 0.060656, 0.020601, 0.065298, 0.066471, 0.062522, 0.048565, 0.027398, 0.106371, 0.019863, 0.025632, 0.022159, 0.021959, 0.11041, 0.037629, 0.073531, 0.100535, 0.026062, 0.06069, 0.044719, 0.043898, 0.023427, 0.020412, 0.026338, 0.04069, 0.030964, 0.02778, 0.042767, 0.172203, 0.028481, 0.024745, 0.040187, 0.029138, 0.041442, 0.041002, 0.045875, 0.028755, 0.026308, 0.02068, 0.023645, 0.027156, 0.03304, 0.029016, 0.0325, 0.024732, 0.031032, 0.035247, 0.023828, 0.042888, 0.03505, 0.033883, 0.026612, 0.02213, 0.02885, 0.020713, 0.021898, 0.027056, 0.053679, 0.05922, 0.039539, 0.048149, 0.0361, 0.036533, 0.212864, 0.035248, 0.023579, 0.04439, 0.045876, 0.039046, 0.027565, 0.048583, 0.025583, 0.044485, 0.058919, 0.040979, 0.043865, 0.031708, 0.023641, 0.034722, 0.026042, 0.024471, 0.022079, 0.073784, 0.028736, 0.030447, 0.025153, 0.024264, 0.069878, 0.052599, 0.029094, 0.035287, 0.07383, 0.042608, 0.05047, 0.487157, 0.153299, 0.044328, 0.114168, 0.163764, 0.050892, 0.022094, 0.041936, 0.089457, 0.024647, 0.038054, 0.040026, 0.02862, 0.03301, 0.114497, 0.061627, 0.057134, 0.023855, 0.09326, 0.024463, 0.030524, 0.033598, 0.023445, 0.026289, 0.072926, 0.068072, 0.043471, 0.167482, 0.036597, 0.050358, 0.05479, 0.054269, 0.039355, 0.028755, 0.048526, 0.096041, 0.040627, 0.04956, 0.195863, 0.021842, 0.052355, 0.036995, 0.030437, 0.076702, 0.036492, 0.089899, 0.135957, 0.038521, 0.036132, 0.047738, 0.025534, 0.036957, 0.080103, 0.023855, 0.041653, 0.032847, 0.052113, 0.062481, 0.050166, 0.017321, 0.028271, 0.032847, 0.02912, 0.07489, 0.042585, 0.024483, 0.02253, 0.037898, 0.030999, 0.032364, 0.035856, 0.023028, 0.026538, 0.022515,

0.039388, 0.030595, 0.07718, 0.067624, 0.02763, 0.044206, 0.034628, 0.033399, 0.030684, 0.036235, 0.034684, 0.024687, 0.024418, 0.037289, 0.02236, 0.068063, 0.045356, 0.044294, 0.021809, 0.022412, 0.080371, 0.064874, 0.066753, 0.044671, 0.045823, 0.055783, 0.046208, 0.09809, 0.046111, 0.060717, 0.031961, 0.065587, 0.028211, 0.024427, 0.042489, 0.043812, 0.071431, 0.054794, 0.038327, 0.091908, 0.36079, 0.10279, 0.040969, 0.022461, 0.037583, 0.012504, 0.069805, 0.067619, 0.033667, 0.144191, 0.04111, 0.020728, 0.027507, 0.027746, 0.036614, 0.031372, 0.031158, 0.022266, 0.041386, 0.028552, 0.042526, 0.033994, 0.043005, 0.035424, 0.074923, 0.028458, 0.033491, 0.049957, 0.089614, 0.052703, 0.02067, 0.106388, 0.035444, 0.015203, 0.054337, 0.028303, 0.024233, 0.173608, 0.022464, 0.029881, 0.02167, 0.043622, 0.04023, 0.031453, 0.024278, 0.036182, 0.052975, 0.056445, 0.048823, 0.043042, 0.082401, 0.02613, 0.027143, 0.04447, 0.055519, 0.031766, 0.031723, 0.037673, 0.009028, 0.036405, 0.04034, 0.022846, 0.022211, 0.037769, 0.027655, 0.050984, 0.036922, 0.025252, 0.029526, 0.18879, 0.032843, 0.025516, 0.020473, 0.035652, 0.064049, 0.023383, 0.020227, 0.026124, 0.025444, 0.021136, 0.030131, 0.049689, 0.057297, 0.183135, 0.036983, 0.04744, 0.03052, 0.03145, 0.027303, 0.025507, 0.071512, 0.047091, 0.025175, 0.109859, 0.058316, 0.037228, 0.095642, 0.040724, 0.041228, 0.034331, 0.128701, 0.050601, 0.069854, 0.19213, 0.035763, 0.021531, 0.039356, 0.094673, 0.027387, 0.137572, 0.020104, 0.044768, 0.040349, 0.114606, 0.032446, 0.066087, 0.029472, 0.089397, 0.038177, 0.099126, 0.068017, 0.033407, 0.041897, 0.097466, 0.020927, 0.024821, 0.039668, 0.058236, 0.050926, 0.034757, 0.027449, 0.08403, 0.02918, 0.026413, 0.035845, 0.066224, 0.031638, 0.036093, 0.035491, 0.074317, 0.024816, 0.02412, 0.023499, 0.034337, 0.10311, 0.061422, 0.032696, 0.039278, 0.074678, 0.037312, 0.043393, 0.03895, 0.085186, 0.022352, 0.049087, 0.024526, 0.046443, 0.031193, 0.023792, 0.043633, 0.033342, 0.0262, 0.293827, 0.037072, 0.08457, 0.03411, 0.022943, 0.043098, 0.023689, 0.020258, 0.028833.